5-2015

Identification and Confirmation of Quantitative Trait Loci for Protein Concentration and Improved Amino Acid Composition in *Glycine max*

Jeneen Samantha Abrams
University of Tennessee - Knoxville, jabrams2@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Jeneen Samantha Abrams entitled "Identification and Confirmation of Quantitative Trait Loci for Protein Concentration and Improved Amino Acid Composition in *Glycine max*." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Plant Sciences.

Vincent Pantalone, Major Professor

We have read this dissertation and recommend its acceptance:

Caula Bey, Dean Kopsell, Arnold Saxton

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Identification and Confirmation of Quantitative Trait Loci for Protein Concentration and Improved Amino Acid Composition in *Glycine max*

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Jeneen Samantha Abrams
May 2015

Acknowledgements

To Teacher Bernard Abrams, you are a loving, supportive and good husband. I feel so blessed to share my life with you and I love you. My daughters: Tabitha, Sarah and Destiny, your love, support and patience means the world to me. You are all intelligent and sweet, Mommy loves you and I anticipate the great things that each of you will do to transform this world. To my father Walter Fields, you have always taught us the value of education and the importance of listening as a part of communication. To my grandmother, Cora Fields, you have taught me how to care for people by the way that you have lived your life, I am so grateful for you. To my sisters Francine and Rah'Shana thank you for your love, belief and support. To the Miller Family and McWilliams Family, thank you for your love and just… everything. To all of my friends, too many to name, love you bunches. To my major advisor, Dr. Vince Pantalone, you are kind and intelligent. You are a diligent and thorough plant breeder and one of the most patient people I have ever known. Thank you for stimulating discussions, thought provoking questions and taking time to help me develop the "breeder's eye". You always made time for me whenever I came to your office or needed your insight in the field. I appreciate you and I will carry that quality with me throughout my career. Thank you for teaching me and placing your confidence in me, I think we did great work together. To my graduate committee, Dr. Beyl, the sweet, strong, and wonderful gentle nudger, there are no words. You are an amazing administrator and an amazing woman. I am thankful to you for so many reasons. Dr. Saxton, you rock!! Thank you for everything you have taught me about statistics, you are one of the greatest statistical minds. I mean really, who writes new SAS macros. I just happened to find out about it on the internet, you are so humble. Dr. Kopsell, thank you for the discussions and asking the hard questions. You have challenged me to grow beyond my comfort zone and I am a better scientist because of you. Thanks to Debbie Ellis and the rest of the Bean Team. To Dr. Catherine Nyinyi Hatcher, thank you for the encouragement and scientific expertise with marker analysis. Dr. Ben Fallen, thank you for your friendship and support. To the farm crew at ETREC, Derick Hopkins, BJ, Vasilj, Brad and Charles, you guys have been my rock. I will never forget everything I have learned from each of you and the way that you went beyond the call of duty to get the work done. Thank you to Dr. Barry Sim, Brad Fisher, Mrs. Rita Hall and the team at HRREC. Thank you Milan team, Dr. Blake Brown, Joseph, Chris B., Allie and Jimmy. Thank you Dr. Jim Orf, Art Killam and the MN team. It has been one of the highlights of this journey to work with such wonderful people. Thank you to the professors, my classmates and administrative staff of the Plant Sciences Dept. A special thank you the Tammy Pass, Angela Berry and Arlene King… you know.

Finally, I dedicate this dissertation to my mother Claudia Ernestine Fields…Mother, I thank you for all the sacrifices you have made so that your daughters could be the best at whatever we chose to do in life. I wish you were here Mommy, but the good seeds that you have sown have produced another doctor. Woman, you were the greatest and I will forever carry your love and tenacity in my heart...The LORD is a sun and a shield. He will give grace and glory. No good thing will He withhold from them who walk upright". *Psalm 84:11.*

Abstract

The demand for poultry and swine production has grown continually as the world population increases. There will be an estimated 8 billion people to feed in the world by 2025. Soybean meal is the most important plant protein ingredient for poultry and swine feedstock. Globally, the US produces 32% of the world's soybean. Although soymeal is nearly a complete plant protein, it is highly supplemented with sulfur-containing amino acids such as cysteine and methionine. The objectives of this project were to utilize molecular markers known as single nucleotide polymorphisms (SNPs) to identify genomic regions associated with protein concentration and amino acid composition. A total of 302 recombinant inbred lines (RIL) were developed from a cross between Essex and Williams 82. $F_{5:8}$ [8^{th} generation of filial generation 5 derived plants] and $F_{5:11}$ [11^{th} generation of filial generation 5 derived plants] lines were used to identify, verify and confirm quantitative trait loci (QTL) associated with protein concentration and amino acid composition. RILs were genotyped with >50,000 SNPs and 17,232 were polymorphic. A total of 9 seed protein QTLs were detected; the QTLs explained 3.1% to 9.8% of variation in seed protein. There were nine seed oil QTL detected which explained 3.2% to 14.1% of the variation in seed oil. No yield QTL were detected. A total of 32 seed amino acid QTLs were detected which explained 4.5% to 14.3% in seed amino acid composition. One seed protein QTL was confirmed on Gm 7. We propose the gene symbol cqSeed Protein-004. Four seed protein QTLs and three seed oil QTLs were positionally confirmed in this study.

KEYWORDS: amino acids, recombinant inbred lines, quantitative trait loci

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction and Literature Review

Introduction

Soybean, *Glycine max* (L.) [Merr.] is one of the most valuable and important agronomic crops grown in the world. Soybean is a member of the Fabaceae family, the third largest land plant family. Fabaceae also includes familiar genera such as peas (*Pisum sativum*), alfalfa (*Medicago sativa*), cowpea (*Vigna unguiculata*) and legume trees such as honey locust (*Gleditsia triacanthos*). The genus *Glycine* is divided into two sub-genera which include domesticated soybean, *Glycine max* (L.) [Merr] and wild soybean, *Glycine soja* Sieb & Zucc. Soybean is indigenous to East Asia and was originally domesticated in China around 1100 BC. Soybean was introduced into the United States (US) in 1765 (Hymowitz, 2004). Soybean was first grown in the US in the early eighteen hundreds as a forage crop, as a coffee (*Coffea arabica*) substitute for soldiers during the Civil War and to produce soy sauce. In 1904, George Washington Carver discovered that soybean was a valuable source of protein and oil (ncsoy.org). Carver also observed an improvement in overall soil health after soybean was grown. He recommended a rotation of cotton (*Gossypium hirsutum*), sweet potatoes (*Ipomoea batatas* (L.) Lam.) and soybean to regional farmers.

Soybean Crop

Soybean is a self-pollinated legume, with the ability to fix atmospheric nitrogen (N_2) utilizing a symbiotic process with soil bacteria called rhizobia. In leguminous plants, the root structures form nodules in which the rhizobia live. The symbiotic relationship provides an energy source for the rhizobia; in return, the rhizobia capture atmospheric nitrogen (N_2). The rhizobia convert N_2 into ammonia (NH_3) and then NH_3 is converted into ammonium (NH_4). The root nodules provide the growing soybean plant with a source of nitrogen, which is essential for

amino acid production. The remains of the harvested soybean plant are left in the soil and they provide a rich and pure source of nitrogen that is released back into the soil. In the soil, the nitrogen contained within the amino acids is released and converted into nitrate (NO_3); the NO_3 is then available to serve as a usable nitrogen source for the plants. In crop rotation, growing soybean is frequently incorporated as a field management practice.

Soybean is photo-sensitive and has been classified as a short-day plant. Short day plants will begin to flower when nights become longer than a critical day length. In the US, soybean is categorized by maturity group (MG). Soybean maturity is determined by the number of days the soybean plants require to reach maturation from planting to harvest (Wilson, 2004). In the US, commercial cultivar maturity zones range from Group 00-VIII (Figure 1.1). Soybean will typically produce a good crop within a distance of 241 km north and south of its maturity zone band. North of the band, the cultivar will flower and mature later than desirable. If soybean is grown south of the band, the same cultivar will mature earlier with reduced stature. Tennessee farmers typically grow MG IV and MG V soybeans. Apical growth of soybean is categorized as determinate or indeterminate. Determinate soybeans complete their vegetative growth cycle prior to flowering. Determinate varieties have bushy plant architecture and are prominent in southern US production. Indeterminate soybeans continue to grow and increase in height even after flowering has occurred. These cultivars tend to be taller and erect and are often grown throughout the mid-west, north central and mid-south regions of the US. In soybean, flowering occurs in response to photosensitivity and soybean maturity group.

Soybean seeds are typically planted after the soil reaches an acceptable temperature of 10°C, however, some soybean can emerge at temperatures as low as 6°C (Lee and Herbeck, 2011). Soybean growth and development occurs during the vegetative and reproductive stages.

During the vegetative stages, structures such as stems, leaves and roots are formed. During the reproductive stages, flowering, seed and pod development, and seed and pod maturation occurs (Fehr, 1977). The vegetative stage is divided into multiple sub-stages and is denoted by (V). During the vegetative emergence (VE) stage, the elongation of the hypocotyl causes the cotyledon to emerge from the soil. The cotyledon is visible above the soil surface within 5-15 days from planting. During the vegetative cotyledon (VC) stage, the cotyledons are clearly visible and unifoliate leaves are fully expanded to provide photosynthesis (7-10 days after emergence). During the V1 stage, the first true trifoliate leaves occur. During the V2 stage, the second trifoliate leaves unroll and active nitrogen fixation begins at the nodules. Additional V stages occur with the formation of each additional trifoliate node and continue for the life of the plant. The soybean reproductive stages are denoted by R. The R1 stage begins when the first blooms become visible. The R2 stage occurs when plants are at full bloom; flower color field notes can be taken when 95% of the plant row has bloomed. The R3 stage is signified by the beginning of pod formation, while surrounding flowers tend to be withered. By the R3 stage, breeders will have already taken field notes for some diseases, however best ratings are taken at peak incidence, which can occur at stages R4 and R5 for some diseases. In the R4 stage, rapid pod growth and the beginning of seed development occurs. The soybean pods are approximately 2 cm in size. In the R5 stage, seeds begin to set and rapid seed filling occurs. Seed set is a term used to signify the visible presence of the developing seed. Both of the R4 and R5 stages are critical to soybean yield. In the R6 stage, full seed development occurs; physiologically the seeds fill the capacity of the pod. In the R6 stage, edamame is also harvested for production. Edamame is a boiled preparation of immature soybean pods; it is often served as cuisine in Asian countries such as Japan and China and has become a popular dish in the US. The R7 stage

denotes the beginning of pod maturity and soybean plants begin to shed their leaves. The R8 stage is denoted by 95% of the pods reaching their mature color. The R8 stage signals to producers that the soybean plants are nearly ready for harvest.

Literature Review

Agronomic Traits

Soybean Seed Characteristics

Typical soybean seeds are composed of approximately 40% protein, 20% oil, 35% carbohydrates and 5% ash (Wilson, 2004). The protein fraction of soybean consists of several smaller protein subunits. Each subunit makes an important contribution to protein concentration and influences the amino acid composition. Soybean protein contains essential and non-essential amino acids, however, soymeal is deficient in the sulfur containing amino acids: cysteine and methionine (Panthee et al., 2006a). Soybean oil is composed of approximately 10% palmitic acid (16:0), 4% stearic acid (18:0), 22% oleic acid (18:1), 54% linoleic acid (18:2) and 10% linolenic acid (18:3) (Wilson, 2004). Soybean seed color can vary in shade from yellow seed coat color to black seed coat color. The major effects on seed coat color are due to the I locus, the R locus and pleiotropic effects of the T locus (Palmer and Kilen, 2004). The dominant I allele affects the anthocyanin production pathway by inhibiting pigment accumulation; the result is the yellow seed coat color (Lindstrom and Vodkin, 1991). The homozygous dominant R locus produces black seeds and homozygous recessive (r) locus produces brown seeds. The T locus is associated with pubescence color and can have a pleiotropic effect on seed-coat color. Plants with tawny pubescence combined with various combinations of the I or R loci can produce black or brown seeds; soybean plants with gray pubescence can produce buff or off-black seeds.

Soybean Agronomic Value

The US is the leading producer of soybean in the world. US soybean growers experienced an outstanding year in 2012, with average sales of $525 per metric ton. US crop revenues for soybean in 2012 were over $43 billion. In 2012, for the first time in US history, soybean and corn hectares were nearly the same (ASA, soygrowers.com). The production of soybean in the US has increased from 66.7 million metric tons in 2003 to 89.5 million metric tons in 2013 (NASS.gov). In 2013, the US produced 32% of the soybean grown in the world. US soybean production is closely followed by Brazil, which produced 31% of the soybean globally. In addition, other top-producing countries included Argentina (19%), China (4%) and India (4%) (Table 1.1). The United States Department of Agriculture estimates world soybean production for 2014/2015 will be 312.8 million metric tons. While soybean is an important crop to the US agricultural economy, is the most important crop produced in the state of Tennessee (TN). In 2013, 1.9 million metric tons of soybean were grown in TN; the estimated crop value of soybean produced in the state was over $682 million (NASS.gov). Tennessee is ranked 15th among the 31 states in the US that produce soybean (NASS.gov).

The value of soybean as an agronomic crop can be attributed to the protein meal and oil produced by the seed. It is one of various oilseed crops from which meal and oil are extracted. In 2013, total world oilseed protein production consisted of 56% soybean, 14% rapeseed (*Brassica napus*), 9% cottonseed (*Gossypium hirsutum*), 8% peanut (*Arachis hypogea*), 8% sunflower (*Helianthus annuus*) and 5% other. Soybean seeds are harvested and transported to processing facilities where they are crushed to extract soybean oil. Soybean meal remains after the crushing process and the meal and oil are used separately to make various products. Globally, soybean is the primary source of plant-based protein meal for animal feedstock (Liu,

1997). Approximately 90% of crushed soybean meal is used to make animal feed (Banaszkiewicz, 2011). In the US, soybean meal accounts for approximately 92% of the total oilseed meal used to feed livestock (Sleper, 2006). In 2013, soybean meal consumption by agriculture animals in the US reached 26.5 million metric tons and 43.0 million metric tons was exported to other countries. Soymeal is used primarily as feed for chickens (poultry) (*Gallus gallus domesticus),* pigs (swine)*(Sus domesticus)* and bovines (*Bos indicus, Bos taurus*). In, 2013, 50% of soybean meal was used in poultry feed, 26% in swine feed, 8% in cattle feed, 11% in dairy feed and 5% in pet foods, fish (*Antiginia spp.*) feed, and miscellaneous feeds (NASS.gov). Soybeans are planted throughout the Midwest, and northern and southern regions of the US. The high nutrient content and adaptability of soybean makes it a good protein source for livestock feed.

Soybean Products

Soybean Meal

Soymeal is a nearly ideal source of nutrition for livestock because it contains a high concentration of essential nutrients. Compared to other vegetable proteins, soymeal contains the highest level of crude protein and provides the most abundant amounts of essential amino acids needed for poultry and swine development (Nielson et al., 1989). The strong potential of soymeal as a feed ingredient is the primary driving force behind soybean market growth (fosfa.org). US soymeal has undergone extensive testing to qualify as the best soymeal in the world, having superior amino acid composition and the best uniformity among batches (soymeal.org). Soybean meal quality is affected by many factors including the balance of protein, fiber and oligosaccharides content. As crude protein increases, the amount of fiber and sugar declines resulting in a higher quality soybean meal (Lindemann and Jang, 2014). Soybean

meal quality can also be affected by its processing method. If soybean is overheated during processing, protein and amino acid quality can become degraded. There are several chemical methods available to evaluate overheating and to determine soybean meal quality. These include urease assays, potassium hydroxide solubility testing and protein dispersibility indices. High values for solubility and dispersability of protein and lower values of urease are desirable characteristics (Lindemann and Jang, 2014).

Soybean meal quality has also been evaluated based on country of origin and meal nutrient content. Mateos et al. (2012) conducted a five year research study to evaluate the nutrient content of soybean meal from the three top producing countries (Table 1.1). After analyzing samples collected from the US, Brazil and Argentina, researchers concluded that US soymeal was higher in crude protein (Table1.2). US soymeal also had a higher sucrose, phosphorus and lysine content than soymeal from Brazil and Argentina. Brazilian soymeal was higher in fiber and iron content. Fiber content in soymeal usually decreases that digestibility of the meal. A recent study of the effects of dietary fiber on nutrient digestibility in pigs revealed that total tract digestibility decreased when total dietary fiber increased (Zhang et al., 2013). Additional studies have been conducted to evaluate feed performance of swine and chickens fed soymeal from major producing countries. A study was conducted to evaluate the growth performance, nutrient digestibility and meat quality for growing-finishing swine fed soymeal from the US, Brazil and India. Swine fed US soymeal were heavier ($p< 0.05$) and also had a greater average daily gain (ADG) and gain per feed ratio (G:F) throughout an 18 week period (Wang et al., 2011). Meat quality among swine was similar among all countries involved in the study. Soymeal from the US, Brazil and India was evaluated to determine broiler and layer performance. Broilers feed US soymeal had a higher G:F than those fed soymeal produced in

Brazil and India, demonstrating that a higher net margin can be accomplished by feeding US soymeal (Park et al., 2002). Layers fed US soymeal had the strongest egg shells and highest egg production; however, treatments were not significantly different for egg production. More recently, an investigation of six different soymeals from South America, US and Spain found that US soymeal had the highest digestibility value (82.3%) compared to other countries included in the study (De Coca-Sinova et al., 2008). Nutrient requirements for livestock do vary, however soybean meal meets most of the requirements for amino acid balance and protein digestibility. The National Research Council has established ideal amino acid composition limits. According to NRC guidelines, commercial soybean meets the majority of the nutritional requirements; however, it is limited in cysteine and methionine. Soybean is currently supplemented with synthetic amino acids to compensate for limited amino acids (NRC, 2014). Additionally, the feed industry has asked the soybean production industry to focus on enhancing lysine, threonine and tryptophan (USB, 2012). While soybean meal has proven to be a valuable source of revenue, soybean oil is also a valuable commodity.

Soybean Oil

Soybean oil is the primary cooking oil utilized globally (Liu, 1997). It is used to produce familiar products such as margarine and shortening. Soybean oil undergoes a stabilization process (hydrogenation) to prevent rancidity. Unfortunately, the current process of hydrogenation converts a portion of the typical soybean oil into *trans* fatty acids, which are unhealthy for long term human consumption and can increase the risk of atherosclerosis. A research study conducted on diet intervention and heart health, found that serum cholesterol levels fell 17% in the intervention group over a one year period (Willett, 2012). The intervention group consumed lower amounts of *trans* fatty acids than the control group. The study concluded

that *trans* fatty acids from partially hydrogenated vegetable oils have clear adverse effects on coronary artery disease and should be eliminated (Willett, 2012). Such evidence poses a threat to the soybean oil industry, but also brings a tremendous opportunity for breeders to develop improved cultivars. Competition with other oil crops such as canola (*Brassica napus*) and sunflower (*Helianthus annuus*) has reduced the demand for soybean oil. Breeders have an opportunity to improve soybean by improving the oleic acid content and reducing linoleic acid for improved fatty acid composition. Such improvements can potentially increase the soybean oil market share.

Soybean oil can also be used to make biodiesel. Biodiesel production has become increasingly more important to the US. In 2007, President Bush signed the Energy Independence and Security Act (EISA) (Pub L.110-140, 2007), which was implemented to increase the production of clean renewable fuel in the US. In Sept 2012, The U.S. Environmental Protection Agency announced an increase in biodiesel volume requirements under the Renewable Fuels Standard (RFS2.) from 1 billion gallons in 2012 to 1.28 billion gallons in 2013 (ASA, 2014). Over 50% of the biodiesel produced in the US comes from soybean. Soybean biodiesel provided 93% more energy than it consumes during production (Hill et al., 2006). The same study also found that soybean biodiesel reduced fossil fuel emissions by 41%, validating the positive environmental impact of soy-based biofuel.

Soybean Products and Derivatives

In addition to animal feed, soybean is an important source of protein for the human diet. Soybean is utilized as a non-meat protein source and soybean meal is used to produce an array of edible products. In 2012, soy-based meatless products accounted for approximately $5 billion in

sales (SFD, 2013). Soy can be converted into textured proteins to become meat alternatives to chicken, turkey, sausage, beef and shrimp. Boca, a large producer of soy-based food, particularly "soy meat", had estimated sales of $3.5 million in 2012. Other soy products include tofu, soymilk, soy energy bars, soy cheese, soy yogurt, and soy frozen desserts. The soy food industry continues to grow and is expected to top $5.2 billion annual sales by 2015.

Soybean derivatives are also heavily utilized in the chemical and baking industries. A surfactant is defined as a product that lowers the surface tension between a liquid and a solid or between two liquids. Soy surfactants are used in many commercial capacities such as in herbicides, detergents, cosmetics, foods and drinks. There are three natural surfactants found in soybean: soy lecithin, soy protein and soy saponin (Oleszek and Hamed, 2010). Egg yolk is the most pure source of lecithin, however, it is too expensive to use consistently in the industrial production of edible goods. Soybean lecithin is available in greater abundance and for a lower cost. Because soybean lecithin has ideal structural and chemical properties, it is widely used as an emulsifier, lubricant, stabilizer, nutrient supplement and wetting agent (Xu et al., 2011). Soybean surfactants are used in carpet fiber production and they are also an important component of paint and ink. Recently, the surfactant market has experienced more strict regulation to protect human health. Soybean produces a natural, renewable surfactant and provides an environmentally safe alternative for surfactant producers.

The entire soybean plant can be used by industry in some capacity. Soybean hulls are used to create biocomposites that can be used to replace building materials that were traditionally made of lumber. Synthetic soybean wood can be used to produce flooring, furniture, countertops and car upholstery materials. In addition, synthetic soybean woods are environmentally friendly because they are biodegradable (Quirino and Larock, 2008).

Seed Components

Seed Protein Concentration

Improving the protein concentration of soybean has been an industry goal for numerous years. The inverse relationship between seed protein and oil has made the pursuit of this goal quite challenging. In addition, increases in protein concentration often adversely affect yield. A delicate balance must be achieved as breeding efforts are directed toward protein trait improvement. To assist breeders in this effort, a table to calculate the theoretical percentage of protein and oil needed to produce 48% soymeal has been developed (Galloway, personal communication). Breeders must achieve improvement of overall protein concentration without significant decreases in soybean oil content (Krishnan et al., 2007). Soymeal that contains 48% protein is designated as High Protein Soymeal and is valued by the industry. Greater protein concentration allows suppliers to use less soymeal to meet appropriate feed protein requirements. Increasing the protein concentration does not affect soybean digestibility, because feed preparations are based on the known protein concentration and are adjusted to maximize digestibility.

To obtain a better understanding of protein concentration, the subunits of soybean protein have also been identified and studied. There are 3 major protein fractions in soybean protein: 11S (Glycinin), 7S (β–Conglycinin) and 2S (α–Conglycinin). The 11S and 7S fractions are the major units that impact protein concentration. The proportion of 11S and 7S fractions present in soybean determine the amount of seed protein manufactured. Both subunits have a direct effect on the quality and quantity of seed protein. The content, ratios and dynamics of biosynthesis of the 11S and 7S fractions may vary with cultivar and environment. The 11S and 7S subunits comprise approximately 70% of soybean total seed protein (Pantalone, 2012). The 11S fraction

contains six acid and basic subunits and accounts for 35% of seed storage protein. The 7S fraction is composed of three subunits: α, α' and β subunits. The 7S fraction has higher solubility and functionality as an emulsifier and provides stability to the 11S fraction (Chove, 2007). Research has focused on the 11S protein fraction because it contains 3-4 times more cysteine and methionine than the 7S fraction (Panthee et al., 2006b). In addition, there is an inverse relationship between the 11S subunit and the 7S subunit. Therefore, soymeal that contains a higher ratio of 11S:7S should also contain elevated cysteine and methionine. Analysis of the two major fractions determined that high protein lines have a higher amount of Glycinin and β –conglycinin (Yaklich, 2001). There are genetic resources available to potentially improve the 11S:7S ratio. For instance, *Glycine soja*, the wild relative of *Glycine max*, could be a useful resource to make improvements in protein quality by improving amino acid composition. If the 11S:7S ratios were adjusted in soybean using *Glycine soja*, it could potentially lead to higher methionine and cysteine composition (Kwanyuen et al. 1997).

The 2S fraction, which is associated with trypsin inhibitors, accounts for 20% of seed protein (Pantalone, 2012). In addition, the 2S subunit accounts for half of the sulfur-containing amino acids in soymeal. Although it has not been investigated in depth, understanding the functionality of the 2S unit may provide further insight on methods to achieve incremental increases in sulfur amino acid composition.

To enhance protein concentration in cultivars, breeders have used various plant introductions (PI). Several soybean PIs are known to contain up to 51% protein (Kim et al., 1996). Classical plant breeding methods rely on phenotypic data to identify traits and develop genotypes with desirable characteristics. These methods have been quite successful in improving numerous traits in soybean, including protein. Since a negative correlation typically

exists between seed yield and protein concentration, it has been challenging to increase protein without significant reduction in yield (Jin et al., 2010). Protein enhancement efforts over the past two decades have included numerous research groups. Additionally, various research populations have yielded information about the relationship between seed components and increased our understanding of protein concentration in soybean. A recurrent selection approach was used to develop 'Prolina,' a high-protein cultivar (Burton et al., 1999). Prolina is a bulk of two F_8-derived lines selected from the first cycle of recurrent selection following the mating of 10 high-protein lines. In Uniform Preliminary Tests, Prolina averaged 461 g kg^{-1} protein and 198 g kg^{-1} oil seed concentration. Its seed protein was 9% higher than 'Centennial', the control, (referred to as checks in soybean experiments) and comparable for oil (200 g kg^{-1}). However, this gain in seed protein for Prolina came at a cost as it yielded 13% less than the check in Uniform Preliminary Tests. In North Carolina Official Variety Trials, Prolina was equal to Centennial for yield.

TN03-350 and TN04-5321 are germplasm lines with improved protein concentration and quality. These lines were released by the University of Tennessee in 2006 (Panthee and Pantalone, 2006a). TN03-350 contained the highest protein concentration among 48 entries in the 2003 maturity Group V Southern Regional Uniform Test, however, yield performance in the Southern Regional Uniform Test was slightly lower than in the check cultivars. TN03-350 performed above the yield checks in independent trials conducted for two years following the 2003 Southern Regional Uniform Test. TN04-5321 was released as a germplasm line because it combined higher protein, increased sulfur containing amino acids and favorable seed yield (Panthee and Pantalone, 2006a). Soybean researchers have continued research efforts to improve soybean protein concentration.

Seed Amino Acid Composition

The animal nutrition and feed formulation industries have charged breeders to develop soybean cultivars with enhanced amino acid profiles. Over 800 million metric tons of synthetic supplements are used annually to fortify soybean with cysteine and methionine alone (Han and Lee, 2000). Notable improvements in soybean amino acid profiles will improve soymeal quality and decrease production costs. There are 20 amino acids; however, soybean breeders are most concerned with essential amino acids. Essential amino acids are those that cannot be produced by the animal, therefore, they must be provided through dietary methods. Essential amino acid requirements differ depending on the animal consuming the feed. Recently five amino acids have been identified by the National Association of Animal Nutritionists as the most important for poultry and swine feed: methionine, cysteine, lysine, tryptophan and threonine.

Amino acids are the structural units that form proteins and they are also involved in key metabolic processes. Methionine helps to metabolize fat and functions as an antioxidant. It is essential for protein synthesis and DNA methylation (a protective mechanism against mutagenesis) (Cavuoto and Fenech, 2012). Cysteine is important for protein biosynthesis, particularly biotin production. Animals have the ability to convert methionine into cysteine; however, each amino acid is needed for specific functions and must be supplied in adequate quantities in the diet (Nikiforova et al., 2002). Lysine is essential for the production of elastin and collagen, both of which are structural proteins for muscles and skin (Radwanski and Last, 1995). Lysine is also important in the production of neurotransmitters and calcium absorption and its deficiency can result in stunted growth, a weakened nervous system and poor immunity in animals (Smirga et al., 2000). Tryptophan is needed for protein synthesis and serves as a precursor for the neurotransmitter serotonin and the vitamin nicotinic acid, which helps the body

digest food and creates energy from carbohydrates, fat, and proteins. Additionally, tryptophan helps to regulate food intake and circadian rhythm (Radwanski and Last, 1995). Threonine is necessary to support cardiovascular, liver, central nervous and immune system function. It is a precursor of glycine and serine, both of which are necessary to produce muscle tissue, elastin and collagen (Lehninger et al., 2000). Of the five amino acids mentioned, cysteine and methionine have occupied most of the current research efforts; they are the sulfur containing amino acids and are heavily supplemented due to their deficiencies in soymeal. The protein portions that help produce sulfur containing amino acids have also been identified as the 11S-glycinin, 7S-β–conglycinin and 2S subunits (Delwiche et al., 2007). Information regarding the genetic regulation of glycinin production had been reported; five genes regulating glycinin protein have been identified (Nielsen et al., 1989). Harada et al. (1989) identified several regions of the soybean genome governing production of 7S; however, genetic x environmental interaction also affects protein concentration and final soybean seed protein composition. As research is conducted to learn more about soybean, heritability, the ability of an organism to genetically transfer the information responsible for the expression of a phenotype, must be considered. Environment has a significant effect on phenotype, therefore stable and heritable traits must be identified. Among USDA soybean seed germplasm accessions, there are numerous soybean accessions with different protein concentrations and amino acid compositions. Additionally, there are well known high protein allele sources such as the Gm 20 allele from Danbaekkong that can be investigated further for gene discovery (Kim et al., 1996). Because such genomic diversity exists, we can be hopeful to find a method to achieve overall protein and amino acid improvement in soybean.

Genetic Background

Molecular Markers

Molecular markers are DNA, RNA or protein based indicators that create a detectable signature within the genome and are associated with qualitative and quantitative traits. Genomic technology is constantly evolving, therefore several foundational analysis techniques have become obsolete. Although the current study utilizes single nucleotide polymorphisms (SNP), several of the techniques used in the past to acquire genomic information remain relevant to our knowledge of soybean seed trait research. DNA markers such a simple sequence repeats (SSR) and single nucleotide polymorphisms, are all based on identification of polymorphisms. Polymorphisms are changes within a genomic sequence which may or may not affect trait expression. On the molecular level, they are the distinguishing factor that helps us to identify genetic differences among individuals. These differences can occur at the allelic or gene level. Changes in sequence, the number of tandem repeats of nucleotides or a single nucleotide change can create significant changes in phenotypic expression. Genomic regions that are identified by the presence of repetitive bases of 1-5 bases (tandem repeats) are called SSRs (Diwan et al., 1997). These regions tend to have highly conserved flanking regions. Polymerase chain reaction (PCR) primers are DNA fragments of a known sequence that form a complementary pairing with DNA strands which have been separated by denaturation. Primers anneal to the flanking regions and changes in the number or length of the tandem repeats can be detected. SSRs are visualized using gel electrophoresis. SNPs are genetic markers that can be identified when a single base change occurs within a DNA sequence. SNPs can be generated by an addition, deletion or substitution in the DNA base pair (Iqbal and Lightfoot, 2005). If a situation occurred where approximately 99.9 % of the DNA between individuals was identical, SNP markers could

distinguish the 0.1% difference between the individuals (Syvanen, 2001). Changes in an individual nucleotide of a DNA sequence can produce linked SNPs or causative SNPs. These changes can occur as single nucleotide additions, deletions or substitutions. Linked SNPs are those that reside outside of the gene and do not affect protein function or expression. Causative SNPs are those that occur within a regulatory or coding region and affect protein function or amino acid expression. Linked SNPs are the primary type of SNP utilized in the current research study. SNP markers have been used with success in soybean research to identify regions of DNA that are associated with quantitative trait loci QTL and to identify additive effects. SNPs are effective because they are the most prevalent form of genetic variation found within the genome and they can be detected at a specific locus when there is an allelic change (Zhu et al. 2003). Using SNPs, more dense genetic maps can be developed and QTLs can be identified more precisely.

Quantitative Trait Loci

A quantitative trait locus is a region that may contain or may be linked to genes associated with a quantitative trait. Quantitative traits are polygenic with each gene generally contributing a small effect. QTL are distinguished utilizing polymorphic molecular markers (SSRs, and SNPs). Such markers are then compared to known phenotypic trait values in a process referred as QTL analysis/mapping. The aim of QTL mapping is to identify regions of the genome that are contributing variation to the trait of interest (Broman, 2001). In QTL analysis, a statistical association is established between phenotypic and genotypic data to identify the approximate location of the QTL within the genome. In a population, a researcher can locate the loci of interest within specific chromosomal regions, estimate the size of their effects, and determine whether their gene action is additive or dominant (Palmer et al., 2004). In soybean,

recombinant inbred lines (RILs) can be utilized for QTL research. RILs are formed by repeated generations of self-pollination in a cross between two parental inbred lines. RILs are useful because such lines tend to offer genetic constituents that are fixed thus enabling accurate genotype evaluation in replicated trials (Broman, 2009). Two strategies can be employed to identify QTL, with the strategy utilized dependent on the type of trait that is being pursued. QTLs can appear in two forms, those that are major are associated with few genes and account for significant variation in genotypes; those that are minor and are associated with many genes and account for small amounts of variation. One of the successful examples of the "few major loci" strategy was its use for soybean cyst nematode (SCN) resistance. SCN resistance was derived from the 'Peking' cultivar and various other plant introductions (Concibido et al., 2004). Molecular markers for resistance have been identified and are linked to a few major loci that account for SCN resistance in soybean. In the first strategy, the molecular markers were identified and used to introduce SCN resistance into elite germplasm (Cahill and Schmidt, 2004). When a quantitative trait results from many genes with a small effect, the first strategy is generally ineffective because the ability to select and secure multiple QTL in a single line decreases as the number of QTL increases. To increase the probability of obtaining the desired genotype, breeders have moved to a strategy of increasing the frequency of the favorable marker alleles in the population to increase the probability of obtaining the desired genotypes. If a breeder develops a mapping population of 100-150 progenies, uses accurate phenotypic data, uses genotypes with markers spaced 10-15 cM apart, and uses an appropriate statistical software package, the process will likely lead to the identification of QTL (Bernardo, 2008)

The soybean genome is approximately 1.1 gigabases (gb) and contains 20 chromosomes. Various forms of DNA markers have been utilized to elucidate information within the soybean

genome. One of the earliest genetic linkage maps of soybean was constructed using restriction fragment length polymorphisms (Keim et al., 1990). The map utilized an F_2 study population of *G. max* x *G. soja* to detect nine quantitative traits including leaf width, leaf length and pod maturity. The map consisted of 26 linkage groups and was mapped at 1200 cM. Only genomic regions responsible for large portions of trait variation could be detected due to the small sample size (60 genotypes). Successive genomic maps were developed and genomic technology has undergone several changes in recent years. More recently discovered molecular tools are being utilized to identify DNA polymorphisms. In 2010, the genome for soybean was sequenced using the shotgun method and scaffolding (Schmutz et al., 2010). The reference sequence for the entire genome was Williams 82, a parent in the current research study. The scaffolds were assembled using SSR and SNP markers. The same markers were also used for accuracy and 46,430 protein coding loci were identified in the soybean genome (Schmutz et al., 2010). Several genetic linkage maps are available for assessing current markers and for genetic alignment of new markers (Shoemaker and Olsen, 1999; Cregan et al., 1999; Song et al., 2004). Song et al. (2004) mapped soybean using 3 populations: *G. max x G. soja* F_2, Minsoy *x* Noir 1 RIL and 'Clark *x* Harosoy' F_2 populations and a total of 606 polymorphic SSRs were identified. A SNP study was also conducted using 'Minsoy *x* Noir 1', 'Minsoy *x* Archer' and 'Evans *x* Peking' (Hyten et al., 2008). Among the populations, 342 SNPs were successfully mapped and 256 of them were polymorphic and were mapped with data from Choi et al. (2007). The results added 256 new markers to the current integrated linkage map. The markers identified were consistent across all populations included in the study and were a valuable refinement of the first consensus map of soybean. Additionally, Song et al. (2013) utilized 52,041 SNPs ("50K SNP") for analysis of University of Tennessee Essex and Williams 82 population. The analysis

produced 17,232 polymorphic SNP loci that were used for genomic analysis in the current research study population. Utilizing molecular information gained from the 50K SNP analysis, we expect to confirm QTL previously reported by other researchers and locate new QTL for protein concentration and amino acid composition in the Essex *x* Williams 82 50K SNP ("ExW8250K") population.

Marker Assisted Selection

For over 50 years, breeding efforts in the US have been focused on increasing crop productivity by increasing yield in various agricultural crops. Breeders have selected for traits such as yield, disease resistance, plant growth habits and form, while nutrient composition has often been ignored (Grusak and Dellapenna, 1999). Marker assisted selection (MAS) is a method by which by breeders utilize genetic markers associated with traits to aid the process of phenotypic selections. MAS enables the selection of superior lines for the trait of interest. Sleper (2006) outlined three specific steps that are required for marker assisted selection: *i*) developing a genomic linkage map, *ii*) identifying the location of molecular markers that co-segregate for the phenotypic trait of interest and, *iii*) selection of plants having molecular markers that are linked to QTL during the breeding process. Marker assisted selection has been used successfully to make selections among soybean populations for various traits. SCN resistance was the first trait in soybean that was widely selected for using MAS. The SCN resistance trait is controlled by a few major QTL and the nature of the trait enables QTL to be utilized effectively to screen for SCN resistance among lines. Cregan et al. (1999) mapped the *rhg1* (first resistance gene for *Heterodera glycines*) locus to *Glycine max* (Gm) chromosome18. The three genes located at the *rhg1* locus are important sources of resistance in soybean from various genetic backgrounds. The *rhg1* locus controls up to 86% of variability for SCN

resistance. Other important QTL such as *rhg2* and *rhg3* have, which provide resistance to race 2 and race 3 soybean cyst nematodes, have also been identified (Guo et al., 2006). Utilization of QTL for MAS for SCN nematode has enabled the development of various resistant soybean lines. Nematode resistance is an important trait due to the high number of soybean plants destroyed by nematodes resulting in> $1 billion is US crop loss annually (Arelli et al., 2015) MAS has been utilized to identify QTL associated with high and low yield among soybean yield trials (Neus, 2010). Lines were grown in 2008 and tested with markers to identify QTL associated with seed yield. Treatment groups were divided into high and low seed yield phenotypes, high and low yield genotypes, and random. The groups were planted the following year in five locations. The genotypic selection method was able to successfully identify lines that would not have been selected due to poor yield performance in 2008 (Neus, 2010). MAS has been utilized to identify genomic regions controlling essential and non-essential amino acids in 282 RIL of Essex x Williams 82. Using the Universal Soy Linkage Panel (USLP) 1.0 of 1,536 SNPs, ten QTL associated with amino acid composition, were identified. The QTL explained 5-14% of the total phenotypic variation for particular amino acids (Fallen et al., 2013). Such studies demonstrate progress toward identifying QTL that are effective for MAS to improve quantitative traits. Therefore, the objective of this study was to focus on detecting, verifying and confirming QTLs associated with protein concentration and amino acid composition to develop a baseline for marker assisted selection within the Essex *x* Williams 82 RIL population.

References

American Soybean Association. 1997. Soy Stats: A Reference Guide to Important Soybean Facts and Figures. http://www.ag.uiuc.edu/»stratsoy/2013/soystats/ (accessed 2 Dec 2014).

Arelli, P.R., V.R. Pantalone, F.L. Allen, A. Mengistu and L.A. Fritz. 2015 Registration of JTN-5203 Soybean Germplasm with Resistance to Multiple Cyst Nematode Populations,J. Plant Reg. 9:108–114.

Banuszkiewicz, T. 2011. Nutritional value of soybean meal, soybean and nutrition, Haney El-Shimmy (Ed.). http://www.intechopen.com/books/soybean-and-nutrition/nutritional-value-of-soybean-meal. (accessed 16 Aug 2013).

Broman, K., 2001. Review of statistical methods for QTL mapping in experimental crosses. Lab Animal 30:44-52.

Broman, K.W. and S. Sen. 2009. A guide to QTL mapping with R/qtl. Springer, New York.

Burton, J.W., T.E. Carter, and R.F. Wilson. 1999. Registration of 'Prolina' soybean. Crop Sci. 39:220- 227.

Cahill, D.J. and D.H. Schmidt. 2004. Use of marker assisted selection in a product development breeding program. Proc. 4th Int. Crop Sci. Cong., 26 Sep.–1 Oct. 2004, Brisbane, Australia. http://www.cropscience.org.au/icsc2004; (accessed 10 Dec 2014).

Cavouto, P. and M.F. Fenech. 2012. A review of methionine dependency and the role of methionine restriction in cancer growth control and life-span extension. Cancer Treatment Rev. 38: 726-736.

Choi I.Y., D.L. Hyten, L. Matukumalli, Q. Song, J. Chaky, C.V. Quigley, K. Chase, K.G. Lark, R.S. Reiter M.S.Yoon, E.Y. Hwang, S.I. Yi, N.D. Young, R.C. Shoemaker, C.P. Van Tassel, J.E. Specht and P.B. Cregan. 2007. A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. Genetics 176:685–696.

Chove, B.E., A.S. Grandison, and M.J. Lewis. 2007. Emulsifying properties of soy protein isolates obtained by microfiltration Food Hydr. 21:8-267-272.

Concibido, V.C., B.W. Diers, and P.R. Arelli. 2004. A decade of QTL mapping for cyst nematode resistance in soybean. Crop Sci. 44:1121–1131.

Cregan, P.B., T. Jarvik, A.L. Bush, R.C. Shoemaker, K.G. Lark, A.L. Kahler, N. Kaya, T.T. VanToai, D.G. Lohnes, J. Chung, J.E. Specht. 1999. An integrated genetic linkage map of the soybean genome. Crop Sci. 39:1464–1490.

De Coca-Sinova, A., D.G .Valencia, E. Jiménez-Moreno, R. Lázaro, and G.G. Mateos, 2008. Apparent ileal digestibility of energy, nitrogen, and amino acids of soybean meals of different origin in broilers. Poul. Sci. 87:2613-2623.

Delwiche, S.R., L.O. Pordesimo, D.R. Panthee, and V.R. Pantalone. 2007. Assessing glycinin (11S) and β-conglycinin (7S) fractions of soybean storage protein by near-infrared spectroscopy. J. Amer. Oil Chem. Soc. 84:1107–1115.

Diwan, N., A.A. Bhagwat, G.B. Bauchan and P.B. Cregan. 1997. Simple sequence repeat DNA markers in alfalfa and perennial and annual *Medicago* species. Genome 40:887-895.

Fallen, B.D., C.N. Hatcher, F.L. Allen, D.A. Kopsell, A.A. Saxton, P. Chen, S. Kantartzi, P.B. Cregan, D.L. Hyten and V.R. Pantalone. 2013. Soybean seed amino acid content QTL detected using the universal soy linkage panel 1.0 with 1536 SNPs. J. of Plant Gen. Sci. 1:68-79.

Federation of Oils, Seeds and Fats Associations (www.fosfa.org) (accessed 16 Dec 2014)

Fehr, W.R. and C.E. Caviness. 1977. Stages of soybean development. Cooperative Extension Service, Agriculture and Home Economics Experiment Station, Iowa State University, Ames, Iowa.

Grusak, M.A., and D. Dellapenna. 1999. Improving nutrient composition in plants to enhance human health. Annu. Rev. Plant Physiol. 50:133-161.

Guo, B., D.A. Sleper, H.T. Nguyen, P.R. Arelli and J.G. Shannon. 2006. Quantitative trait loci underlying resistance to three soybean cyst nematode populations in soybean PI 404198A. Crop Sci. 46:224-233.

Han I. and J.H. Lee. 2000. The role of synthetic amino acid in monogastric animal production. Asian-Aus J. Ani. Sci. 13:543-560.

Harada, J.J., S.J. Barker, and R.B. Goldberg. 1989. Soybean B-conglycinin genes are clustered in several DNA regions and are regulated by transcriptional and posttranscriptional processes, The Plant Cell 4:415–425.

Hill, J., E. Nelson, D. Tilman, S. Polasky, and D. Tiffany**.** 2006. Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. Proc. Natl. Acad. Sci. U. S. A. 103:11206 –11210.

Hymowitz, T., Speciation and Cytogenetics. 2004. In: Boerma, H.R. and Specht, J.E. (Eds.), Soybeans: Improvement, Production, and Uses, Third Edition. ASA, CSSA, SSSA, Madison, WI, USA, p. 621–677.

Hyten, D.L., Q. Song, Y. Zhu, I.Y. Choi, M-S. Yoon, J.E. Specht, L.K. Matukumalli, R.L. Nelson, R.C. Shoemaker, N.D. Young, P.B. Cregan. 2008. High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor. Appl. Genet. 116: 945-952.

Iqbal, M.J., and D.A. Lightfoot. 2005. Application of DNA markers: soybean improvement. Molecular marker systems in plant breeding and crop improvement. Springer Berlin Heidelberg Publishing. Carbondale, IL. pp. 371-385.

Jin J., X. Liu, G. Wang, L. Mi, Z. Shen, X. Chen, S.J. Herbert. 2010. Agronomic physiological contributions to the yield improvement of soybean cultivars released from 1950 to 2006 in Northeast China. Field Crops Res. 11:6-123.

Keim, P., B.W. Diers, T.C. Olson, and R.C. Shoemaker. 1990. RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics 126:735-742.

Kim, S.D., E.H. Hong, Y.H. Kim, S.H. Lee, Y.K. Seong, K.Y. Park, Y.H. Lee, Y.H. Hwang, E.H. Park, H.S. Kim, Y.H. Ryu, R.K. Park, and Y.S. Kim. 1996. A new high protein and good seed quality soybean cultivar Danbaekkong. J of Agr Sci. 38:1228-1232.

Krishnan, H., S.S. Natarajan, A.A. Mahmoud and R.L. Nelson. 2007. Identification of glycinin and beta conglycinin subunits that contribute to the increased protein concentration of high-protein soybean lines. J Agric Food Chem 55:1839-1845.

Kwanyuen, P., V.R. Pantalone, J.W. Burton and R.F. Wilson. 1997. A new approach to genetic alteration of soybean protein composition and quality. J of the Amer Oil Chem Soc 74:983-987.

Lee, C. and J. Herbeck. 2011. Soybean planting in Kentucky. Cooperative Extension Service. Pub. AGR-130.

Lehninger, A.L., D.L. Nelson, and M.M. Cox. 2000. Principles of Biochemistry 3rd ed. W. H. Freeman, New York.

Lindemann, M. and Y.D. Jang. 2014. The Value of U. S. Soybean Meal for Pigs. University of Kentucky. USA

Lindstrom J.T. and L.O. Vodkin 1991. A soybean cell wall protein is affected by seed color genotype. Plant Cell 3:560-571.

Liu, K., 1997. Chemistry and Nutritional Value of Soybean Components, Soybean: Chemistry, Technology and Utilization, Aspen Publishers, Gaithersburg. p. 25-35.

Mateos, G.G., M.P. Serrano, P. Guzman and P. Lazaro. 2012. Evaluating the nutrient value of soybean meals according to origin: an opportunity to the feed industry. USSEC. The Nutritional Value of US Soybean Meal http://www.soymeal.org/ReviewPapers/NutritionalValueUSSoybeanMealWhitePaper.pdf (accessed 15 Dec 2014)

NASS - National Agricultural Statistics Service. www.nass.usda.gov. (accessed 22 Aug 2013)

North Carolina Producers Association. www.ncsoy.org (accessed 22 Dec 2014)

National Research Council. www.nas/edu/nrc. (accessed 22 Dec 2014)

Nielsen, N.C., C.D. Dickinson, T.J. Cho, V.H. Then, B.J. Scallon, R.L. Fischer, T.L. Sims, G.N. Drews, and R.B. Goldberg. 1989. Characterization of the glycinin gene family in soybean. The Plant Cell 1:313–328.

Neus. J. 2010. Marker assisted selection for seed yield in soybean [Glycine max (L.) Merr.] plant row yield trials. PhD Dissertation. University of Illinois, Champaign, IL.

Nikiforova, V., S. Kempa, M. Zah, S. Maimann, O. Kreft, A.P. Casazza, K. Riedel, E. Tauberger, R. Hoefgen and H. Hesse. 2002. Engineering of cysteine and methionine biosynthesis in potato. Amino Acids 22:259-278.

Oleszek, W. and A. Hamed. 2010. Saponin-based surfactants, In Surfactants from renewable resources, Wiley Series in Renewable Resources, Kjellin, M. & Johansson, I. (Ed.), 239-252, Wiley, West Sussex.

Pantalone, V.R. 2012. Modern Breeding Approaches for Enhancing Soybean Protein Quality, pp. 189-218. In R. F. Wilson (ed) Designing Soybeans for 21st Century Markets.

Panthee, D.R., V.R. Pantalone, A.M. Saxton, D.R. West, C.E. Sams, J.H. Orf and A.S. Killam. 2006a. Quantitative trait loci controlling sulfur containing amino acids, methionine and cysteine, in soybean seeds. Theor. Appl. Gen. 112:546-553.

Panthee, D.R. and V.R. Pantalone. 2006b. Registration of soybean germplasm lines TN03-350 and TN04-5321 with improved protein concentration and quality. Crop Sci. 46:2328-2329.

Palmer, R.G. and T.C. Kilen. 1987. Qualitative genetics and cytogenetics. In L. R Wilcox (ed.) soybeans: Improvement, production and uses. Second ed. Agronomy 16:135-210.

Palmer, R.G., T.W. Pfeiffer, G. R. Buss, and T. C. Kilen. 2004. Qualitative Genetics. In: Boerma, H.R., and Specht, J.E. (Eds.), Soybeans: Improvement, Production, and Uses, Third Edition. ASA, CSSA, SSSA, Madison, WI, USA, pp. 137-233.

Park, Y.A., H.K. Kim, H.S. Lee, I.S. Shin and K.Y. Whang. 2002. Effect of three different soybean meal sources on layer and broiler performance. Asian Astral J An Sci 15:254-265. Public Law 110-140 2007. Energy Independence and Security Act of 2007. http://www.gpo.gov/fdsys/pkg/PLAW-110publ140/pdf/PLAW-110publ140.pdf

Quirino, R. L. and Larock R.C. 2008. Synthesis and properties of soy hull reinforced biocomposites from conjugated soybean. J Appl. Poly Sci. 107:423-430.

Radwanski, E.R., and R.L. Last. 1995. Tryptophan biosynthesis and metabolism; Biochemical and molecular genetics. The Plant Cell. 7:921-934.

Schmutz J., S.B. Cannon, J. Schlueter, J. Ma., T. Mitros, W. Nelson, D.L Hyten., Q. Song, J.J. Thelen, J. Cheng, et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463:178-183.

Shoemaker, R.C., and T.C. Olson. 1993. Molecular linkage map of soybean (Glycine max L. Merr). Genetic maps: locus maps of complex genomes. Cold Spring Harbor Laboratory Press, New York. pp.131-138.

Sleper, D.A., J.M. Poehlam. 2006. Breeding Field Crops. Blackwell Publishing. Ames, IA. p 130-133.

Smriga, M., Mori, M., and K.Torii. 2000. Circadian release of hypothalamic norepinephrine in rats in vivo is depressed during early L-lysine deficiency. J. Nutr 130:1641-1643.

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Fickus, R.L. Nelson and P.B. Cregan. 2013 "Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean," PLoS One 8:1 Article ID:e54985

.

Song, Q.J., L.F. Marek, R.C. Shoemaker, K.G. Lark, V.C. Concibido, X. Delannay, J.E. Specht, P.B. Cregan. 2004. A new integrated genetic linkage map of the soybean. Theor. Appl. Genet. 109:122–128.

Soy Foods Directory. http://www.soyfoods.com/ (accessed 8 Sept 2013)

Soybean Information Center. http:// www.soymeal.org. (accessed 5 May 2014).

Soypheno Maturity date calculator, Mississippi State University http://msucares.com/deltasoy/gsModelFrame.htm (accessed Sept 3, 2013).

Syvanen, A.C. 2001. Accessing genetic variation: Genotyping single nucleotide polymorphisms. Genet. 2:930-942.

United Soybean Board. http:// www.unitedsoybean.org. (accessed 8 Sept 2014)

Wang, J.P., S.M. Hong, L.Yan, J.H. Cho, H.S. Lee and I.H. Kim. 2011. The evaluation of soybean meals from 3 major soybean-producing countries on productive performance and feeding value of pig diets. J. of An Sci 89:2768-2773.

Willett, W.C. 2012. Dietary fats and coronary heart disease. J. Intern Med 272: 13–24.

Wilson, R.F. 2004. Seed composition. In: Boerma, H.R. and Specht, J.E. (Eds.),Soybeans:Improvement, production, and uses, 3rd Edition. ASA, CSSA, SSSA, Madison, WI, USA, pp. 621–677.

World Soybean Production 2014/2015. http:// www.worldsoybean production.com/ ?gclid=CKPk8Tzz8ICFchr7AodCkIAtg.. (accessed 17 Dec 2014)

Xu, Q., Nakajima, M., Liu, Z. and T. Shiina. 2011. Soybean-based Surfactants and Their Applications, Soybean - Applications and Technology, Prof. Tzi-Bun Ng (Ed.), ISBN: 978-953-307-207-4, InTech.

Yaklich, R.W. 2001. Beta-conglycinin and glycinin in high-protein soybean seeds. J. Agric. Food Chem. 49:2 729–735.

Zhang, W., D. Li, L. Liu, J. Zang, W. Yang and L. Zhang. 2013. The effects of dietary fiber level on nutrient digestibility in growing pigs. J. Ani. Sci and Biotech. 4:17.

Zhu, Y.L., Q.J. Song, D.L. Hyten, C.P. van Tassell, L.K. Matukumalli, D.R. Grimm, S.M. Hyatt, E.W. Fickus, N.D. Young, and P.B. Cregan. 2003. Single-nucleotide polymorphisms in soybean. Genet. 163:1123–1134.

Appendix 1: Chapter 1 Tables and Figures

Table 1.1 Value of global soybean production in the top five producing countries in 2013[a]

Country	Production (Million Metric Tons)	Percentage of Global Soybean Production
United States	89.5	32
Brazil	87.5	31
Argentina	54.0	19
China	12.2	4
India	11.0	4

[a] Soystats 2013

Table 1.2 Percentage of protein in soymeal from the top three producing countries in 2007-2012[a]

Country	% Crude Protein	# of samples
United States	47.3	164
Brazil	46.6	131
Argentina	45.4	136

[a] Mateos et al. 2012

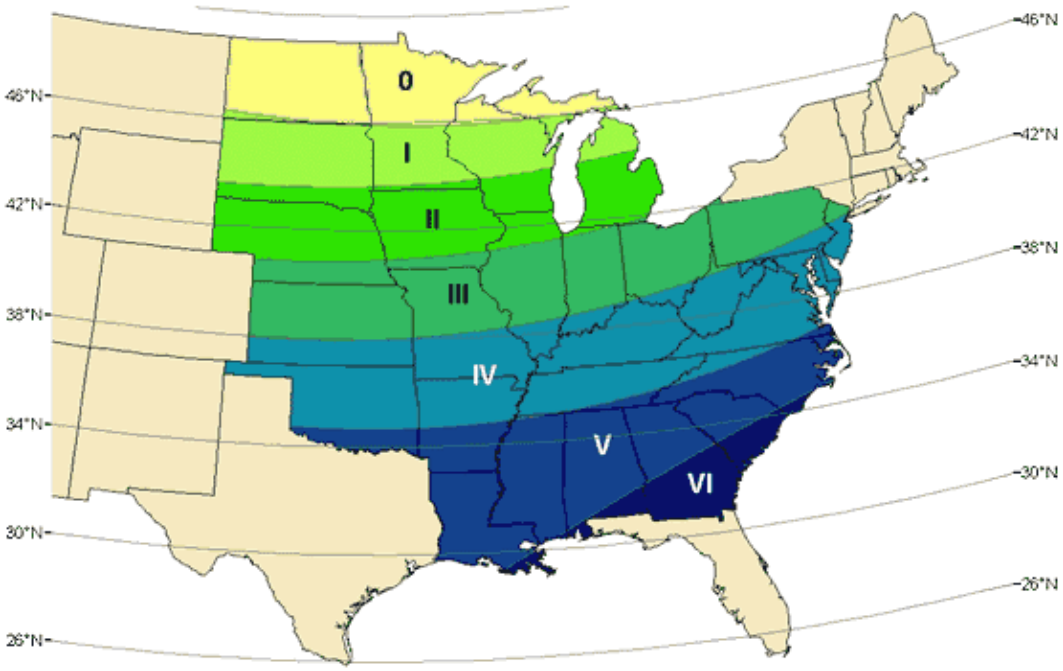

Figure 1.1 Soybean maturity zones with latitude and maturity groupings
Courtesy of fullpotentialoutdoors.com

.

Chapter 2: Identification and Confirmation of QTL for Seed protein, Seed Oil and Yield in a Essex *x* Williams 82 RIL population

J. S. Abrams, B.D. Fallen, C. N. Hatcher, A.M. Saxton, C.A Beyl, D.A. Kopsell, J.H. Orf , A.S. Killam, P. Cregan, D. Hyten and V.R. Pantalone

Abstract

Soybean is an excellent source of plant protein. Soymeal is used as a primary component of animal feed and the oil produced by the seed is used in cooking and for biodiesel. Soybean yield is always a focus of a modern breeding program, however seed quality traits such as seed protein concentration and seed oil content must also be improved. The objectives of this study were to identify, verify and confirm quantitative trait loci (QTL) for seed protein, seed oil, and yield for the 'Essex *x* Williams 82' recombinant inbred line (RIL) population. For the current study $F_{5:8}$ RILs were phenotyped as progeny rows in one environment and genotypic data from a >50,000 SNP marker analysis were used to identify initial QTL. A total of 17,232 SNP markers were polymorphic. A linkage map was constructed using the mapping population and QTL were detected using composite interval mapping. The $F_{5:11}$ population was phenotyped across three environments for seed protein, seed oil and yield. Genotypic data were used to verify previously reported QTLs and to identify additional seed protein, seed oil and yield QTLs. Protein and oil had high heritabilities across multiple environments, however yield heritability was low. Seed protein and seed oil were negatively correlated. Yield was only mildly affected by seed protein concentration with a weak negative correlation (-0.15). Genotype and genotype x environment interaction were significant ($p< 0.05$) for seed protein, seed oil and yield. Based on testing in one environment during Year 1, four seed protein and two seed oil QTL were detected. Seed protein QTL explained 3.1-4.5% of seed protein and seed oil QTL explained 4.4%-4.8% of the variation in seed oil content. No QTL for yield were detected. In 2013, recombinant inbred lines were tested over three environments. Five seed protein QTLs were detected. One seed protein QTL was confirmed on Gm 7. Seed protein QTLs explained 3.1%-9.8% of variation in seed protein. Seven seed oil QTLs were detected seed oil QTLs explained 3.2%-14.1% of the variation in oil content. No yield QTLs were detected in 2013.

KEYWORDS: quantitative trait loci, recombinant inbred lines, composite interval mapping

Introduction

The United States (US) is the leading global producer of soybean [*Glycine max* (L.) Merr.] (Soystats, 2014). Soy protein is highly valued as livestock feed because it produces a rich soybean meal which contains nutritious protein and multiple amino acids. The protein and amino acid content of soymeal provides essential nutrients for livestock growth and development. Soybean is also used for human consumption as a protein and meat substitute. Soybean provides approximately 54% of the vegetable oil consumed in the US (Soystats, 2014). Usage of soybean oil as biofuel has increased steadily over the past decade due to the federal mandate of the Energy Independence and Security Act (2007). Soybean accessions can vary in protein and oil content, however, soybean seeds typically contain 40% protein and 20% oil on a dry matter basis (DM). Protein and oil improvement has been an important objective, although yield is the primary objective in most breeding programs. Soybean yield can change seed protein concentration because seed protein and yield are negatively correlated (Specht et al., 2001). The inverse relationship between seed protein and yield has had a significant impact on the ability to improve simultaneously seed protein and yield traits in soybean. Marginal improvements in seed protein concentration could improve soymeal nutritional profiles and increase soymeal profitability. In addition, improvement in the seed oil composition and quality would help soybean to maintain its current market share in the seed oil industry and enhance its marketability. Higher yield is important because it ultimately increases profitability for farmers.

Molecular markers such as single nucleotide polymorphisms (SNP) have been used effectively to identify quantitative trait loci (QTLs) associated with seed protein concentration, seed oil content and yield (Panthan et al., 2013; Fallen et al., 2015). Continual identification and confirmation of valuable QTLs will contribute to the overall improvement of soybean seed

quality traits. However, breeders must use the QTLs in breeding programs for marker assisted selection (MAS) to achieve the benefit. As we identify QTLs, it is important to pursue QTLs with relevant pedigrees and those that maintain stability over multiple environments.

Seed Quality QTL

Seed Protein and Oil QTL

Seed protein and seed oil QTLs have been identified on each of the 20 chromosomes in soybean (Diers et al., 1992; Brummer et al., 1997; Orf et al., 1999; Wang, et al., 2004). Diers et al. (1992) documented major QTLs associated with protein and oil on *Glycine max* chromosome (Gm) 15 and 20. Seed protein and seed oil QTLs were detected in two populations "Young *x* PI 416937" and "PI 97100 *x* Coker 237" (Lee et al., 1996), and these were detected in multiple environments. Increased seed protein was associated with decreased seed oil in the PI 97100 x Coker 237 population. Yesudas et al. (2013) studied Essex *x* Forrest recombinant inbred lines (RILs) using composite interval mapping (CIM). Four QTLs associated with seed protein were found on Gm 2, 6, 12 and 20. Essex was the high protein parent and the Essex allele was intrinsic to higher protein at four out of five loci. Two seed oil QTL were also identified in the Essex *x* Forrest population at Gm 18 and Gm 20, the increase in seed oil was attributed to the Forrest allele. Using F_6 -derived RILs developed from a cross of N87-984-16 *x* TN93-99, Panthee et al., (2005) located a significant seed protein QTL on Gm 20 and seed oil QTL on Gm, 1, 10 and 12. The QTL located on Gm 6 and Gm 20 have been of great interest because they have been identified in multiple cultivars and environments (Panthee et al., 2005; Bolon, 2010). Using 216 RILs of Magellan *x* PI 438489B and 156 Magellan *x* PI 567516C RILs, Pathan et al. (2013) identified seven seed protein QTLs and six seed oil QTLs across various environments. Two QTLs common for seed protein and seed oil were detected on Gm 5 and 6. These QTLs

were previously identified in the Magellan *x* PI 438489B and Magellan *x* PI 567516C. However, the QTLs have not been confirmed in Soybase. Confirmation of a soybean QTL requires: a) a separate meiotic event and environment than when QTL was originally mapped, b) at least one parent in common with the original mapped study and the confirmed study and c) an experiment -wise error rate of 0.01 or lower (soybase.org). Confirmed QTLs for protein and oil QTL on Gm 5 and Gm 6 could be important targets to identify candidate genes involved in protein and oil content modifications (Panthan et al., 2013). QTLs that show consistency can be utilized by breeders in marker assisted selection. Hyten et al. (2004) identified four protein QTLs from 131 RILs. Three seed protein QTL were conferred by Essex located on Gm 7, 9 and 13 at 41.9 cM, 114.0 cM, and 15.8 cM, respectively. One seed protein QTL on Gm 6 was conferred by Williams was located at 119.8 cM. The allele on Gm 6 conferred by Williams contributed to high protein concentration, but was also linked to an early maturity allele (Hyten et al., 2004). In the same study, six seed oil QTL were identified on Gm 1, 6, 7, 17 and 19, conferred by the Essex and Williams alleles. Additional seed protein QTLs were identified by screening 176 $F_{2:4}$ RILs of PI97100 *x* Coker 237 R (Fasoula et al., 2004). Three PI97100 *x* Coker 237 seed protein QTL were identified previously and received a confirmed designation as confirmed QTL (cqQTL) in Soybase.

According to Soybase, nearly 150 seed protein, 200 seed oil and 150 yield QTLs have been identified (Soybase, 2015). Only two seed protein QTLs, four seed oil QTLs and one yield QTL have been confirmed. A cqQTL designation demonstrates genomic consistency and cqQTL would likely be utilized for MAS with higher efficacy.

Numerous years of research have been invested to identify seed protein, seed oil and yield QTL. As a scientific community, the lack of QTL confirmation limits the ability to

confidently utilize QTL for MAS, even though unconfirmed QTL are being utilized in some breeding programs. Use of MAS could enable significant progress to be made in producing higher yielding soybean with improved seed quality traits. Our efforts were aimed at moving toward solidifying the location of QTLs associated with the traits of interest. In addition to identifying confirmed QTLs, we sought to verify and positionally confirm QTLs. Verified QTLs are those which have been identified among lines that may differ in generation, however the lines share the same parentage. For example the $F_{5:8}$ and $F_{5:11}$ lines were used in this research study. QTL that are found in both years would constitute a verified QTL. Positionally confirmed QTL are those which are located on the same chromosome in the same or similar position, however they may have different parentage. These types of QTL are beneficial because they tend to occur in highly conserved regions of the genome. Evidence suggests that identification and confirmation of QTL will be a measureable step toward progress in our ability to understand and improve seed protein, seed oil and yield through molecular breeding strategies.

The objectives of this study were 1) identify new QTLs, verify and/or confirm QTL for seed protein, seed oil and yield in the 'Essex *x* Williams 82', 50,000 SNP RIL population (ExW82-50K); 2) to test the stability of seed protein and seed oil QTL across multiple environments; and 3) to positionally confirm QTLs for seed protein, seed oil and yield.

Materials and Methods

Plant Materials

The University of Tennessee ExW82-50K RIL mapping population was used to conduct the current research study. The research population was derived from a cross between Essex 86-15-1 (E) and Williams 82-11-43-1 (W82); the numbers refer to reselections. Essex is a southern cultivar with purple flower color and gray pubescence. It has a determinant growth habit and is

classified as a maturity group V soybean (Smith and Camper, 1973). 'Williams 82' is a cultivar which carries a resistance gene for *Phytopthora* (Bernard and Cremeens, 1988). It was derived from the northern cultivar 'Williams' and has white flowers and tawny pubescence. It has an indeterminate growth habit and is classified as a maturity group III soybean (Bernard and Lindahl, 1972).

The nomenclature Essex 86-15-1 refers to a within line reselection process using the 86th plant, growing the seeds from the row as a single row then choosing the 15th plant, followed by growing seeds as a single row and choosing the 1st plant. The same process was used to form Williams 82-11-43-1. In 2005, the seeds were planted in the crossing block and a genetic cross of the reselected lines designated Essex 86-15-1 *x* Williams 82-11-43-1 was made at the University of Tennessee, East Tennessee Research and Education Center (ETREC). The cross produced Essex *x* Williams 82, which was the F_1 progenitor of the ExW82-50K population. The F_1 seed from the ExW82-50K cross were harvested and grown at the Tropical Agricultural Research Station (TARS) in Isabela, PR. The F_2 population was advanced to the F_5 generation through the single seed descent method (Brim, 1966). The F_2 and F_3 generations were grown at ETREC in 2006 and 2007, respectively. The F_4 plants were grown at the TARS during the spring of 2008. In the summer of 2009, $F_{4:5}$ seeds were planted and became F_5 plants in Beltsville, MD at a USDA greenhouse. Each plant was tagged for identification and leaf tissue was collected for 50K SNP analysis. Seeds were harvested from 1021 individual plants. $F_{5:6}$ seeds were planted in Homestead, FL in fall 2009 for a seed increase. The seeds were harvested and planted in the spring 2010 at ETREC as $F_{5:7}$ rows. $F_{5:8}$ seeds were harvested and sent to Homestead, FL for a seed increase in fall 2010. $F_{5:9}$ seeds were planted for a yield experiment in 2011 and produced $F_{5:10}$ seeds.

Due to discrepancies in USDA-ARS flower color data and 2011 field data, flower colors were verified by growing the lines in the greenhouse at the University of Tennessee in fall 2012. $F_{5:10}$ remnant seeds were sent to Homestead, FL for a seed increase. The $F_{5:11}$ seeds were planted in spring 2013 for the current research study.

Field Methods

Soybean entries from the 2010 study were planted in Knoxville, TN at ETREC. Each line was planted as one rep in a two row plot. Row length was 6 m and rows spacing was 76 cm. In 2013, the initial ExW82-50K mapping population (1021 plants) was subdivided into three categories based on maturity: Maturity Group (MG) III, MG IV and MG V. The MG V population included 302 lines that were selected to conduct this research study. Maturity documentation differed in 2010 and 2013. Additionally three yield checks were utilized [(Osage (Chen et al., 2007), 5002T (Pantalone et al., 2004) and Ellis (Pantalone, 2015)]. Soybean entries for the 2013 study were planted in 6.1 m length two-row plots in a randomized complete block design (RCBD). Research plots were established at three locations with three replications. ETREC is located at Knoxville, TN (35.53°N 83.57°W). The soil at this location was classified as Etowah loam and the annual average rainfall is 1193.8 mm. Highland Rim Research and Education Center (HRREC) is located in Springfield, TN (36.28°N 86.51°W). The soil type at HRREC is classified as either Dickson silt loam or Sango silt loam and the annual average rainfall is 1244.6 mm. The Research and Education Center at Milan (RECM) is located in Milan, TN (35.54° N 88.44° W). The soil is classified as Loring B2 series fine silt and annual average rain fall is 1371.6 mm.

Agronomic Traits Evaluation

In 2010 and 2013, seeds were monitored for germination rates through field observation and all experimental plots were evaluated for agronomic traits. Phenotypic data were collected for flower color, pubescence color, lodging, height and maturity. Flower color was noted as purple, white or segregating when 95% of the plants had bloomed. Plant height and lodging were measured at maturity. Lodging was scored on a 1-5 scale, with 1 representing plants that were upright and 5 representing plants that were prostrate. Maturity was recorded when 95% of the pods had achieved their mature color. Pubescence was scored as gray, tawny or segregating, when 95% of the pods in the plot showed their mature color (Fehr and Caviness, 1977). All plots were rogued to ensure genetic integrity.

For the 2010 study, maturity dates were documented according to the Julian calendar. The first maturity was recorded on day 251 (Sept 8, 2010) and the last maturity was recorded on day 288 (Oct 15, 2010). The RIL lines were placed into MG III through MG V based on maturity dates. In 2013, maturity was documented with Day 1 as September 1. Because all lines were classified as MG V, the first maturity date was recorded on Sept 27th, corresponding to day 27 in the maturity log. The 2010 study dates were converted to a Sept 1 start date for data analysis.

Near Infrared Analysis for Protein Concentration

Phenotypic data for protein concentration and oil content were collected for each research plot. A 25 g whole bean subsample was ground using a water-cooled Knifetec 1095 Sample Mill (FOSS Tecator, S-26321, Hogana, Sweden). The samples were ground for 20 seconds and placed into Whirl Pak bags (Nasco, Fort Atkinson, WI). Samples were barcode labeled using the ZM 1000 Barcode Printer System (Zebra Technologies, Lincolnshire, IL). The ground samples

were analyzed at the Soybean Analysis Lab at the University of Minnesota, St. Paul, MN (Dr. Jim Orf). The laboratory was equipped with a JET Air Filtration System (LaVergne, TN) to minimize particulate matter from the ground soybean samples. A subsample weighing 12.5 g was placed into a small sample cup to conduct the ground bean analysis. The sample was leveled using a spatula and placed into the Perten near infrared (NIR) analyzer (Hagersten, Sweden). The Perten software conducted an initial scan of auto diagnostics for instrument response, wavelength accuracy and NIR repeatability. The scan reported the soybean composition percentages which were converted to g kg^{-1} [per kilogram] units. Several samples were selected for wet chemistry verification to assure accuracy of NIR readings (Table 2.4)

Genotyping

F_5 plants were tagged for identification and DNA was extracted at the Soybean Genomics Laboratory at the USDA Beltsville Agricultural Research Center (USDA-ARS) in Beltsville, MD. Samples containing 50 µl [microliters] of DNA at a 200 ng/µl content. The samples were assayed using the GoldenGate® assay with >50,000 SNP markers following the protocol from the manufacturer and methods described by Hyten et al. (2008) and Fan et al. (2003). All samples were assayed using the Illumina BeadStation 500G (Illumina, San Diego, CA). The population produced 17,232 polymorphic SNP markers. Data from the Beltsville analysis was used for the genetic analysis component of the current research population.

Data Analysis

Analyses of variance (ANOVA) were conducted on the phenotypic data to determine if there were significant differences among the RIL genotypes, environment and genetic x environment (g x e) interaction. All ANOVA were conducted using the MIXED procedure of SAS (SAS ver. 9.3, Cary, NC). ANOVA were conducted to detect significant differences among

RILs for protein concentration, oil content and yield. Initially, all effects were tested as random factors and included genotype, environment, genotype x environment and genotype x replication within environment. To allow greater precision and higher power, replication within environment was removed from the model. The initial model was found to be too conservative and negated a high proportion of measurable variation. The final model was the following:

$$Y_{ij} = \mu + B_i + T_j * + B*T_{ij}$$

where, Y is the observation of the *j*th treatment (genotype) in the *i*th block (environment).

A second analysis was conducted using genotype as a fixed term and environment and genotype x environment as random terms. Fisher's LSD was performed for mean separation with the MIXED procedure of SAS (SAS ver. 9.3, Cary, NC). The least squares means were calculated and compiled to use in QTL analysis.

Pearson's correlation analysis was performed to determine phenotypic correlations among seed protein, seed oil and yield (CORR procedure, SAS ver. 9.3, Cary, NC). To determine the portion of phenotypic variation among RILs that resulted from genetic differences and estimate heritability, restricted maximum likelihood estimation of variance components was used. A broad sense estimate of heritability of seed protein , seed oil and yield in the population was calculated on an entry mean basis (Nyquist, 1991) as follows:

$$h^2 = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_{ge}/e + \sigma^2/re}$$

where, h^2 represents the heritability, σ^2_g is genotypic variance, σ^2_{ge} is genotype x environment variance, σ^2 is error variance, *r* is number of replications and *e* is number of environments.

Since the population was F_5-derived, most of the genetic variance was additive. Therefore, we are obtaining an approximation of narrow sense heritability with this formula.

Linkage Map and Quantitative Trait Loci Analysis

R/qtl (Broman and Sen, 2009) was used to construct a genetic linkage map using 17,232 SNP markers and 302 genotypes. The estimated map length was 2072 cM and utilized 12,730 markers, after unlinked markers were discarded. Chromosomal location, marker order and position were determined by composite interval mapping (CIM) (Broman, 2001; Broman and Sen, 2009). A standard walking speed of 2 cM was used to conduct CIM (Broman and Sen, 2009). Ten thousand permutations were performed to establish a log odds (LOD) threshold of 3.0 at α=0.001 (Müller-Myhsok, 2009). The LOD threshold of 3.0 was consistently applied to identify seed protein , seed oil and yield on each of the twenty soybean chromosomes.

Results

Phenotypic Analysis of Seed protein, Seed Oil and Yield

Significant differences were observed among RIL genotypes for seed protein concentration, seed oil content and yield ($p<0.001$). The differences among RILs for environment were not significant for seed protein($p=0.1594$), oil ($p=0.1592$) or yield ($p=0.1785$), however, genotype and genotype x environment were significant for seed protein, seed oil and yield ($p<0.001$).

A moderate negative correlation was found between seed protein and seed oil among RILs of Essex *x* Williams 82. Our results confirm previous reports of the inverse relationship between seed protein and seed oil in many soybean populations (Burton, 1987;Wilcox and Goudong, 1997; Wilson, 2004; Panthee, 2005). However, yield was only weakly correlated with seed protein (Table 2.1). Seed oil and yield had a weak positive correlation (Table 2.1). Seed

protein concentration for parents and checks ranged from 363.4 g kg^{-1} to 394.2 g kg^{-1} (Table 2.2). Seed protein concentration for the RIL population ranged from 331.6 g kg^{-1} to 461.2 g kg^{-1} crude protein (Table 2.3). The RIL population maximum differed from the mean by 79.0 g kg^{-1} seed protein . Several transgressive segregates that produced higher seed protein than the high parent Essex and all checks (Ellis, 5002T and Osage), were identified (Figure 2.1). Seed oil content for parents and checks ranged from 209.8 g kg^{-1} to 225.4 g kg^{-1} (Table 2.2). The seed oil content among RILs ranged for 193.1 g kg^{-1} to 248.5 g kg^{-1} oil (Table 2.3). Several lines were also identified higher than the high parent for oil, Williams 82 (Figure 2.2). Additionally, several RILs had greater yields higher than the high parent Essex (Figure 2.3) and all checks (Table 2.2). Seed yields ranged from 1872.7 kg ha^{-1} to 5553.3 kg ha^{-1} (Table 2.3). The Essex *x* Williams 82 cross takes advantage of the wide array of variation in traits from crossing cultivars of southern and northern heritage. Heritability estimates were high for seed protein and seed oil with R^2 values of 87.4% and 84.2%, respectively (Table 2.3). However, heritability for yield was low at 52% (Table 2.3).

Quantitative Trait Loci Analysis of Seed Protein, Seed Oil and Yield

'Essex x Williams 82' phenotypic data for 2010 and >50, 000 SNP markers were analyzed using R/ qtl (Figure 2.4). Based on ten thousand permutations and a LOD threshold of 3.0 (Figure 2.5), the analysis revealed four QTLs on Gm 6, 7, 13, and 14. Two seed oil QTL were also found on Gm 6 and Gm 14. LOD scores of the seed protein QTLs ranged from 3.1 to 4.5 and the individual QTLs explained between 7.2% - 9.0% of the variation in seed protein (Table 2.6). LOD scores of seed oil QTL ranged from 4.4 to 4.8 and individual QTL explained 7.1% - 8.6% of the variation in seed oil (Table 2.7). Yield QTLs were not observed.

The QTLs could be used to identify putative top performing lines for seed protein and seed oil based on the presence of favorable allelic combinations (Table 2.7).

The 2013 data produced QTLs for seed protein on Gm 6, 7, 9 and 13 for genotypes grown in Knoxville, TN; Gm 2, 6, 7, 9 and 13 for genotypes grown in Springfield, TN and Gm 6, 7, 9, and 13 for genotypes grown in Milan, TN. Four out of five seed protein QTLs were present in every environment. However, there was a QTL on Gm 2 that appeared in the Springfield location only. The LOD scores for seed protein QTLs ranged from 3.1 to 9.8 and the R^2 value explained 4.4% -11.8% of variation in seed protein (Table 2.6). Seed oil QTLs were identified on Gm 6, 9, 10, 11, 13, 18 and 19. The seed oil QTLs had LOD scores that ranged from 3.6 to14.1 and the R^2 explained 4.1%-8.3% of the variation in seed oil (Table 2.6). No yield QTLs identified in 2013.

Discussion

Phenotypic Analysis of Seed protein, Seed Oil and Yield

Protein and oil are economically important seed quality traits in soybean. In addition, yield is always an important breeding objective. Essex and Williams 82 parents have smaller differences in seed protein and seed oil than lines typically tested. However, there were significant differences among RIL for protein, oil and yield ($p< 0.0001$). Heritability followed trends previously identified in several soybean populations with high heritability for seed protein and seed oil (Chung et al., 2003, Hyten et al., 2004, Panthee et al., 2006, Pathan et al., 2013). Differences among RILs were due to genetic variation. Yield had low heritability in this population and similar findings have been reported in other research studies (Palomenque, 2010; Orf, 1999). However, several lines produced yields that were higher than Osage, the highest yield check, in the study.

Typically, high protein and low protein lines are crossed to form RILs. Minor changes such as those observed among the Essex x Williams 82 RIL population suggest minor genetic components can have a significant impact on protein production. The formation of lines containing sufficient amounts of protein and oil allowed the formation of transgressive segregates which produced higher seed protein and oil than the parents and checks (Table 2.3). Transgressive segregates also produced higher yields than both of the parents and checks (Figure 2.3). Mean protein concentration was 381.8 g kg^{-1}, mean oil content was 224.8 g kg^{-1} and mean yield was 3562.1 kg ha^{-1}. Essex was the high parent for protein and was higher than the mean protein for the RILs. Likewise, Williams 82 was the high parent for oil and was higher than the RILs mean for oil (Table 2.2, Table 2.3). The results identified in this study were similar to previous findings and selections would be effective for achieving genetic gain for protein, oil and yield.

Quantitative Trait Loci Analysis of Seed Protein, Seed Oil and Yield

Seed protein QTLs were identified on Gm 6, 7, and 13 in both years. The QTLs were also present in all environments during the second year. The presence of the QTLs across all environments at these loci, indicate a strong genetic effect. To conduct additional comparisons, information was gathered from the integrated genetic linkage map for the soybean genome and updated the integrated genetic linkage map of the soybean (Cregan et al., 1999: Song et al., 2004). The information was used to form associations with previous genetic markers that have been identified near the QTLs identified in this study. The QTL located on Gm 6 is located near microsattellite marker (Satt) 148, which is also associated with seed yield. The QTL located on Gm 7 (43.3-43.7 cM) was also identified by Hyten et al. (2004) at Gm 7 at 41.9 cM and both conferred by the Essex allele. The Gm 7 seed protein QTL meets the criteria for a confirmed

seed protein QTL in Soybase. We therefore, propose the QTL symbol cq004-Seed protein for the QTL on Gm 7. The QTL is located near the 3' flanking region of Satt 253 and the 5' flanking region of Satt 540, which has also been identified as a seed protein QTL. Additional seed protein QTL were only detected in one environment. There was a seed protein/ seed oil QTL located on Gm 14 in Knoxville location during the first year, and seed protein QTL on Gm 2 during second year at the Springfield location. The presence of Essex allele at Gm 2 locus increased seed protein by 2.4 g kg^{-1}. The absence of these particular QTL in a multi-year test could provide information regarding the effect of environment on expression of alleles associated with seed protein production and may denote epigenetics effects. Genetic x environment interaction has a significant role in phenotypic expression. All QTL reported on a given chromosome were within 10 cM in distance of one another in order to be reported as the same QTL. The distance is reasonable length in which QTL being compared may be considered the same QTL (Broman, 2001). The QTL located on Gm 13 is located near the 3' untranslated region site of mRNA, an area that can govern protein expression (soybase.com). While several QTL were identified, we did not identify any QTLs on Gm 20 (LG I), a chromosome that is well documented as a major QTL for seed protein (Diers et al., 1992; Brummer et al., 1997; Chung, 2003). Several positionally confirmed QTL (pcQTL) were identified for seed protein. Two seed protein pcQTL were located on Gm 7 at 43.7 and 50.0 cM and Gm 9 at 63.0 cM. QTL at similar positions were previously identified by Eskardari et al., 2013. A seed protein pcQTL was also found on Gm 14 at 45.5 cM, which was identified previously by Kabelka et al., 2004 .

Only two seed oil QTLs were identified in year one. A total of seven were identified in year two, of those, 1 was a verified QTL on Gm 6 at 54.0 cM. There were four pc QTL identified for seed oil, one was located on Gm 9 at 6.1 cm was first identified by Qi et al., 2011.

Two pcQTLs on Gm 10 at 3.0 cM and Gm 11 at 21.2 cM were identified by Brummer et al., 1997. An additional pcQTL was identified on Chr 13 at 0.8 cM, which corresponds to the QTL identified by Qi et al., 2011. Although pcQTL show some genetic consistency because they were found among various populations, other findings suggest a significant environmental effect on seed oil content. Seed oil content can be greatly affected by rainfall and temperature (Carrera et al., 2011). Several seed oil QTLs were found in a similar region as QTL affecting seed protein, suggesting that pleiotropic effects may occur among loci affecting seed protein and seed oil content. The seed protein QTL and seed oil QTL on Gm 6 were less than 3 cM apart in distance and a similar situation was observed on Gm 14.

Conclusions

Yield QTL were not detected in this experiment suggesting that a significantly higher number of genotypes are needed to increase experimental power for yield QTL discovery in Group V Essex *x* Williams 82 soybeans lines. While it is important to continue additional studies to identify QTL associated with seed protein, seed oil and yield, we must focus on validation and confirmation of QTL that have been discovered. We successfully detected a QTL on Gm 7 at 50.0 cM and propose this QTL as cq004-Seed protein in Soybase. QTL discovery is helpful to plant breeding schemes only if the QTL are applied in breeding programs (Bernado, 2008). Perhaps larger studies of different parentages are needed to gain a comprehensive view of the impact of QTL identification on seed quality traits and yield in soybean.

References

Bernard R.L. and C.R. Cremeens. 1988. Registration of Williams 82 soybean. Crop Sci. 28:1027–1028.

Bernard, R.L. and D.A. Lindahl. 1972. Registration of Williams soybean. Crop Sci. 12:716.

Bernardo, R. 2008. Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 years. Crop Sci 48:1650-1664.

Bolon, Y.T., B. Joseph, S.B. Cannon, M.A. Graham, B.W. Diers, A.D. Farmer, G.D. May, G.J., J.E., Specht, Z.J. Tu, N. Weeks, W.W. Xu, R.C. Shoemaker and C.P. Vance. 2010. Complementary genetic and genomic approaches help characterize the group I seed protein QTL in soybean. BMC Plant Biol. 1471-2229.

Brim, C.A. 1966. A modified pedigree method of selection in soybeans. Crop Sci. 6:220.

Broman, K. 2001. Review of statistical Methods for QTL mapping in experimental crosses. Lab Animal 30:44-52.

Broman, K.W. and S. Sen. 2009. A guide to QTL mapping with R/qtl. Springer, New York.

Brummer, E.C., G.L. Graef, J.H. Orf, J.R. Wilcox, and R.C. Shoemaker. 1997. Mapping QTL for seed protein and oil content in eight soybean populations. Crop Sci. 37:370–378.

Burton, J.W. 1987. Quantitative genetics: Results relevant to soybean breeding. In: J.R. Wilcox, editor, Soybean: Improvement, production, and uses. 2nd ed. ASA, CSSA, and SSSA, Madison, WI. p. 211–247.

Carrera, C. M.J. Martinez, J. Bardanelli and M. Balzarint. 2011. Environmental variation and correlation of seed components in nontransgenic soybeans: protein, oil,unsaturated fatty acids, tocopherols, and isoflavones. Crop Sci. 51:800-809

Chung, J., H.L. Babka, G.L. Graef, P.E. Staswick, D.J. Lee, P.B. Cregan, R.C. Shoemaker, and J.E. Specht. 2003. The seed protein, oil, and yield QTL on soybean linkage group I. Crop Sci. 43:1053–1067.

Cregan, P.B., T. Jarvik, A.L. Bush, R.C. Shoemaker, K.G. Lark, A.L. Kahler, N. Kaya, T.T. VanToai, D.G. Lohnes, J. Chung, J.E. Specht. 1999. An integrated genetic linkage map of the soybean genome. Crop Sci. 39:1464–1490.

Diers, B.W., P. Keim, W.R. Fehr, and R.C. Shoemaker. 1992. RFLP analysis of soybean seed protein and oil content. Theor. Appl. Genet. 83:608–612.

Energy Independence and Security Act of 2007 (EISA), Pub. L. No. 110-140,121 Stat. 1492, 1783-84 (Dec, 19, 2007), codified at 42 U.S.C. 17381

Eskandari, M., E. R. Cober, and I. Rajcan. 2013. Genetic control of soybean seed oil: QTL and genes that increase oil concentration without decreasing protein or with increased seed yield. Theor. Appl. Genet. 126:1677-1687.

Fan, J.B., A. Oliphant, R. Shen, B.G. Kermani, F. Garcia, K.L. Gunderson, M. Hansen, F. Steemers, S. L. Butler, P. Deloukas, L. Galver, S. Hunt, C. McBride, M. Bibikova, T. Rubano, J. Chen, E. Wickham, D. Doucet, W. Chang, D. Campbell, B. Zhang, S. Kruglyak, D. Bentley, J. Haas, P. Rigault, L. Zhou, J. Stuelpnagel, M.S. Chee. 2003. Highly parallel SNP genotyping. Cold Spring Harb. Symp. Quant. Biol. 68:69–78.

Fasoula,V.A., D.K. Harris, R.H. Boerma. 2004. Validation and designation of quantitative trait loci for seed protein, seed oil, and seed weight from two soybean populations. Crop Sci. 44:1218-1222.

Fehr, W.R. and C.E. Caviness, 1977. Stages of soybean development. Cooperative Extension Service, Agriculture and Home Economics Experiment Station, Iowa State University, Ames, Iowa.

Hyten, D.L., V.R. Pantalone, C.E. Sams, A.M. Saxton, D. Landau-Ellis, T.R. Stefaniak, and M.E. Schmidt. 2004. Seed quality QTL in a prominent soybean population. Theor. Appl. Genet. 109:552–561.

Hyten, D.L., Q. Song, Y. Zhu, I.Y. Choi, M-S. Yoon, J.E. Specht, L.K. Matukumalli, R. L. Nelson, R.C. Shoemaker, N.D. Young, P.B. Cregan. 2008. High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor. Appl. Genet. 116: 945-952.

Kabelka E.A., B.W. Diers, W.R. Fehr, A.R LeRoy, I.C. Baianu, T. You, D.J, Neece, R.L. Nelson. 2004. Putative alleles for increased yield from soybean plant introductions. Crop Sci. 44: 784-791.

Lee, S.H., M.A. Bailey, M.A.R. Mian, T.E. Carter Jr., E.R. Shipe, D.A. Ashley, W.A. Parrott, S. Hussey, and H.R. Boerma. 1996. RFLP loci associated with soybean seed protein and oil content across populations and locations. Theor. Appl. Genet. 93:649–657.

Müller-Myhsok. 2009. Estimating experiment-wise error rate by permutations. mga.bionet.nsc.ru/~yurii/courses/esp29.../permlecture_2009_6spp.pdf 9 (accessed 17 Feb 2015).

Nyquist W.E. 1991. Estimation of heritability and prediction of selection response in plant populations. Crit. Rev. Plant Sci. 10: 235–322.

Orf, J.H., K. Chase, T. Jarvik, L.M. Mansur, P.B. Cregan, F.R. Adler, and K.G. Lark. 1999. Genetics of soybean agronomic traits: I. Comparison of three related recombinant inbred populations. Crop Sci. 39:1642–1651.

Palomeque, L., L.J. Liu, W. Li, B.R. Hedges, E.R. Cober, M.P. Smid, L. Lukens, and I. Rajcan. 2010. Validation of mega-environment universal and specific QTL associated with seed yield and agronomic traits in soybeans. Theor Appl Genet. 120: 997–1003.

Pantalone, V.R. 2012. Modern Breeding Approaches for Enhancing Soybean Protein Quality, pp. 189-218. In R. F. Wilson (ed) Designing Soybeans for 21st Century Markets.

Panthan, S.M., T. Vuong, K. Clark, J.D. Lee, J.G. Shannon, C.A. Roberts, M.R. Ellersieck, J.W. Burton, P.B. Cregan, D. L. Hyten, H. T. Nguyen, and D. A. Sleper. 2013. Genetic mapping and confirmation of quantitative trait loci for seed protein and oil contents and seed weight in soybean. Crop Sci. 53: 765-774.

Panthee, D.R., V.R. Pantalone, D.R. West, A.M. Saxton, and C.E. Sams. 2005. Quantitative trait loci for seed protein and oil content, and seed size in soybean. Crop Sci. 45: 2015-2022.

Qi, Z., Q. Wu, X. Han, Y. Sun, X. Du, and C. Liu. 2011. Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes. Euphytica 179:499-514.

Smith, T.J. and H.M. Camper. 1973. Registration of 'Essex' Soybean. Crop Sci. 13:495-495.

Song, Q.J., L.F. Marek, R.C. Shoemaker, K.G. Lark, V.C. Concibido, X. Delannay, J.E. Specht, and P.B. Cregan. 2004. A new integrated genetic linkage map of the soybean. Theor. Appl. Genet. 109:122–128.

Soybase and Soybean Breeder's Toolbox. http://www.soybase.org/ accessed 14 Jan 2015

Soy Stats 2014. A reference guide to important soybean facts and figures. http://www.soystats.com/ (accessed 5 Jan 2015).

Specht, J.E., M. Germann, J.P. Markwell, K.G. Lark, J.H. Orf, M. Macrander, K. Chase, J. Chung, and J.L. Graef. 2001. Soybean response to water: A QTL analysis of drought tolerance. Crop Sci. 41:493–509.

Wang, D., G.L. Graef, A.M. Procopiuk, and B.W. Diers. 2004. Identification of putative QTL that underlie yield in interspecific soybean backcross populations. Theor. Appl. Genet. 108:458–467.

Wilcox, J.R., and Z. Goudong. 1997. Relationships between seed yield and seed protein in determinate and indeterminate soybean populations. Crop Sci. 37:361–364.

Wilson, R.F. 2004. Seed composition. 2004. In: Boerma, H.R. and Specht, J.E. (Eds.), Soybeans: Improvement, Production, and Uses, Third Edition. ASA, CSSA, SSSA, Madison, WI, USA, pp. 621–677.

Yesudas, C.R., R. Bashir, M.B. Geisler, and D.A. Lightfoot. 2013. Identification of germplasm with stacked QTL underlying seed traits in inbred soybean population from cultivars Essex and Forrest. Mol Breeding 33:3.

Appendix 2: Chapter 2 Tables and Figures

Table 2.1 Pearson's correlation coefficient between protein, oil and yield in 302 $F_{5:11}$-derived recombinant inbred lines of Essex 86-15-1 x Williams 82-11-43-1 grown in 2013 Knoxville, TN; Springfield, TN; and Milan, TN.

Trait	**Protein**†	**Oil**†	**Yield**
Oil †	-0.69**		
Yield (kg ha^{-1})	- 0.15**	0.15**	
Seed Weight	0.05*	0.27**	0.38**

* p=0.05
**p=0.01
†DM, dry matter

Table 2.2 Means of protein, oil and yield of soybean seed for parents and checks grown 2013 in Knoxville, TN; Springfield, TN; and Milan, TN.

	Parent Means		Check Means		
Trait	Essex	Williams 82	Ellis	5002T	Osage
Protein (DM) g kg^{-1} †	393.8	386.5	366.8	363.4	394.2
Oil (DM) g kg^{-1} †	217.1	226.2	213.8	225.4	209.8
Yield kg ha^{-1}	3088.3	2593.6	3805.8	3622.6	3983.1

† DM, dry matter

Table 2.3 Descriptive statistics of protein and oil (g kg^{-1} seed) and yield (kg ha^{-1}) of soybean seed from 302 $F_{5:11}$-derived recombinant inbed lines of Essex 86-15-1 x Williams 82-11-43-1 grown in 2013 in Knoxville, TN; Springfield, TN; and Milan, TN.

Trait	Min	Mean	Max	$LSD_{0.05}$	h^2 (%)
		(g kg^{-1} seed)			
Protein †	331.6	382.2	461.2	24.1	87.4
Oil †	193.1	224	248.5	13.3	87.2
Yield (kg ha^{-1})	1872.7	3095.4	5553.3	1348.2	52.0

† DM dry matter

Table 2.4 Wet lab results of random soybean seed protein samples from 302 $F_{5:11}$-derived recombinant inbred lines of Essex 86-15-1 x Williams 82-11-43-1 grown in 2013 in Knoxville, TN; Springfield, TN; and Milan TN.

Sample ID	Total N % (initial run)	Wet Lab	NIR
94405	5.7	38.1	41.9
93505	6.0	40.4	44.0
92514	6.2	41.7	45.7
94519	5.8	39.0	43.6
93729	6.5	43.3	47.2
92813	5.6	37.9	42.9
92296	5.9	38.9	39.3
92583	5.8	38.9	38.6
92783	5.6	37.4	37.8
92897	5.6	36.8	36.7
93511	5.7	37.9	37.0
93818	6.0	39.6	38.7
94182	6.3	41.5	40.2
94369	6.1	40.4	39.1
94430	6.2	40.8	40.3
94593	6.1	40.6	40.6
94676	5.7	37.7	37.0
94785	5.8	38.9	37.4
	Avg	39.4	40.4

Table 2.5 Quantitative trait loci identified using R/qtl for composite interval mapping located on various chromosomes associated with protein and oil concentration in 302 $F_{5:8}$ derived recombinant inbred lines of Essex 86-15-1 x Williams 82-11-43-1 grown in 2010 in Knoxville, TN.

Location	Trait	QTL Name	Chr†	MLG‡	Molecular Marker	Loc§ (cM)	LOD¶	Confidence Interval of QTL position	R^2 (%)	Effect# g kg^{-1}
Knoxville, TN	Protein	Seed protein 36-1	Gm 6	C2	Gm06_45433980_G_A	57.2	4.5	53.0-61.7	7.3	0.5(W)
Knoxville, TN	Protein	cqSeed protein 004	Gm 7	M	Gm07_14773717_G_T	50.0	3.2	37.1-75.0	7.2	0.4 (E)
Knoxville, TN	Protein	Seed protein - 36-2	Gm 13	F	Gm13_1395656_T_C	199.9	3.1	198.0-204.5	7.6	0.5 (E)
Knoxville, TN	Protein	Seed protein 36-3	Gm 14	B2	Gm14_30024382_T_C	45.2	4.1	42.9-46.6	9.0	0.5 (W)
Knoxville, TN	Oil	Seed Oil 39-1	Gm 6	C2	Gm06_45362447_C_T	55.0	4.4	52.0- 58.0	7.1	0.4 (E)
Knoxville, TN	Oil	Seed Oil 39-2	Gm 14	B2	Gm14_34892670_G_A	45.5	4.8	42.0-51.0	8.6	0.4 (E)

†Chr., chromosome.
‡ MLG=Molecular Linkage Group
§ The QTL position was determined based on genetic linkage map constructed in the present study, measured in centimorgans.
¶ LOD, logarithm of the odds.
# Effect indicates quantitative change in protein and oil concentration associated with either (E) Essex 15-86-1 or (W) Williams 82-11-43-1 allele.

Table 2.6 Quantitative trait loci identified using R/qtl for composite interval mapping located on various chromosomes associated with protein concentration in 302 $F_{5:11}$-derived recombinant inbred lines of Essex 86-15-1 x Williams 82-11-43-1 grown in 2013 in Knoxville, TN; Springfield, TN; and
Milan, TN.

Location	Trait	QTL Name	Chr†	MLG‡	Molecular Marker	Loc § (cM)	LOD ¶	Confidence Interval of QTL position	R^2 (%)	Effect # g kg^{-1}
Springfield, TN	Protein	Seed protein 36-4	Gm 2	D1b	Gm02_11030750_C_T	24.0	4.2	12.0-36.0	4.4	2.4 (E)
Knoxville, TN	Protein	Seed protein 36-5	Gm 6	C2	Gm06_45433980_G_A	57.2	6.3	52.0-61.7	7.7	3.9 (W)
Springfield, TN	Protein	Seed protein 36-6	Gm 6	C2	Gm06_45433980_G_A	57.2	6.3	51.0-60.3	7.6	3.8 (W)
Milan, TN	Protein	Seed protein 36-7	Gm 6	C2	Gm06_47758592_C_T	65.5	3.9	52.0-83.0	6	2.9 (W)
Knoxville, TN	Protein	Seed protein 36-8	Gm 7	M	Gm07_10236359_A_G	43.7	4.2	32.0-66.4	6.2	2.9 (E)
Springfield, TN	Protein	Seed protein 36-9	Gm 7	M	Gm07_14773717_G_T	43.3	3.1	39.2 – 74.0	3.8	2.3 (E)
Milan, TN	Protein	Seed protein 36-10	Gm 7	M	Gm07_18237983_G_A	66.3	3.5	36.0-75.4	5.7	2.6 (E)
Knoxville, TN	Protein	Seed protein 36-11	Gm 9	K	Gm09_38637679_A_C	63.0	5.9	56.5-67.0	7.3	3.2 (W)
Springfield, TN	Protein	Seed protein 36-12	Gm 9	K	Gm09_38385411_A_G	61.1	6.5	56.6-65.6	8.8	3.5 (W)
Milan, TN	Protein	Seed protein 36-13	Gm 9	K	Gm09_38922926_G_A	64.0	4.0	58.0-67.0	6.9	2.9 (W)
Knoxville, TN	Protein	Seed protein 36-14	Gm 13	F	Gm13_36573410_T_G	183.0	9.8	176.0-187.0	11.7	4.0 (E)
Springfield, TN	Protein	Seed protein 36-15	Gm 13	F	Gm13_35370448_C_T	183.0	8.7	180.8-187.7	11.8	4.1 (E)
Milan, TN	Protein	Seed protein 36-16	Gm 13	F	Gm13_36573410_T_G	189.0	4.9	181.8-192.0	7.4	3.0 (E)

† Chr., chromosome.
‡ MLG=Molecular Linkage Group
§ The QTL position was determined based on genetic linkage map constructed in the present study, measured in centimorgans.
¶ LOD, logarithm of the odds.
# Effect indicates quantitative change in protein concentration associated with either (E) Essex 15-86-1 or (W) Williams 82-11-43-1 allele.

Table 2.7 Quantitative trait loci identified using R/qtl for composite interval mapping located on various chromosomes associated with oil concentration in 302 $F_{5:11}$-derived recombinant inbred lines of Essex 86-15-1 x Williams 82-11-43-1 grown in 2013 in Knoxville, TN; Springfield, TN; and Milan, TN.

Location	Trait	QTL Name	Chr†	MLG‡	Molecular Marker	Loc § (cM)	LOD ¶	Confidence Interval of QTL position	R^2 (%)	Effect # g kg^{-1}
Milan, TN	Oil	Seed Oil 39-3	Gm 9	K	Gm09_2563451_T_C	6.1	3.6	0.0-13.5	4.1	1.2 (W)
Milan, TN	Oil	Seed Oil 39-4	Gm 10	O	Gm10_1669760_G_A	3.0	4.1	0.0-20.0	6.4	1.5 (W)
Milan, TN	Oil	Seed Oil 39-5	Gm 11	B1	Gm11_4216279_A_G	21.2	3.7	10.0-23.9	4.9	1.3 (E)
Milan, TN	Oil	Seed Oil 39-6	Gm 13	F	Gm13_211725_G_A	0.8	3.2	0.0-160.3	3.8	1.1 (E)
Milan, TN	Oil	Seed Oil 39-7	Gm 18	G	Gm18_2117841_T_C	5.3	3.3	0.0-67.0	4.1	1.2 (W)
Milan, TN	Oil	Seed Oil 39-8	Gm 19	L	Gm19_42089062_C_T	196.6	14.1	194.8-199.8	6.3	1.5 (W)
Springfield, TN	Oil	Seed Oil 39-9	Gm 6	C2	Gm06_22004394_C_T	54	3.7	51.0-60.0	8.3	2.4 (E)
Springfield, TN	Oil	Seed Oil 39-10	Gm 19	L	Gm19_42089062_C_T	196.6	12.7	194.8-201.6	7.8	3.0 (W)

† Chr., chromosome.
‡ MLG=Molecular Linkage Group
§ The QTL position was determined based on genetic linkage map constructed in the present study, measured in centimorgans.
¶ LOD, logarithm of the odds.
# Effect indicates quantitative change in oil concentration associated with either (E) Essex 15-86-1 or (W) Williams 82-11-43-1 allele.

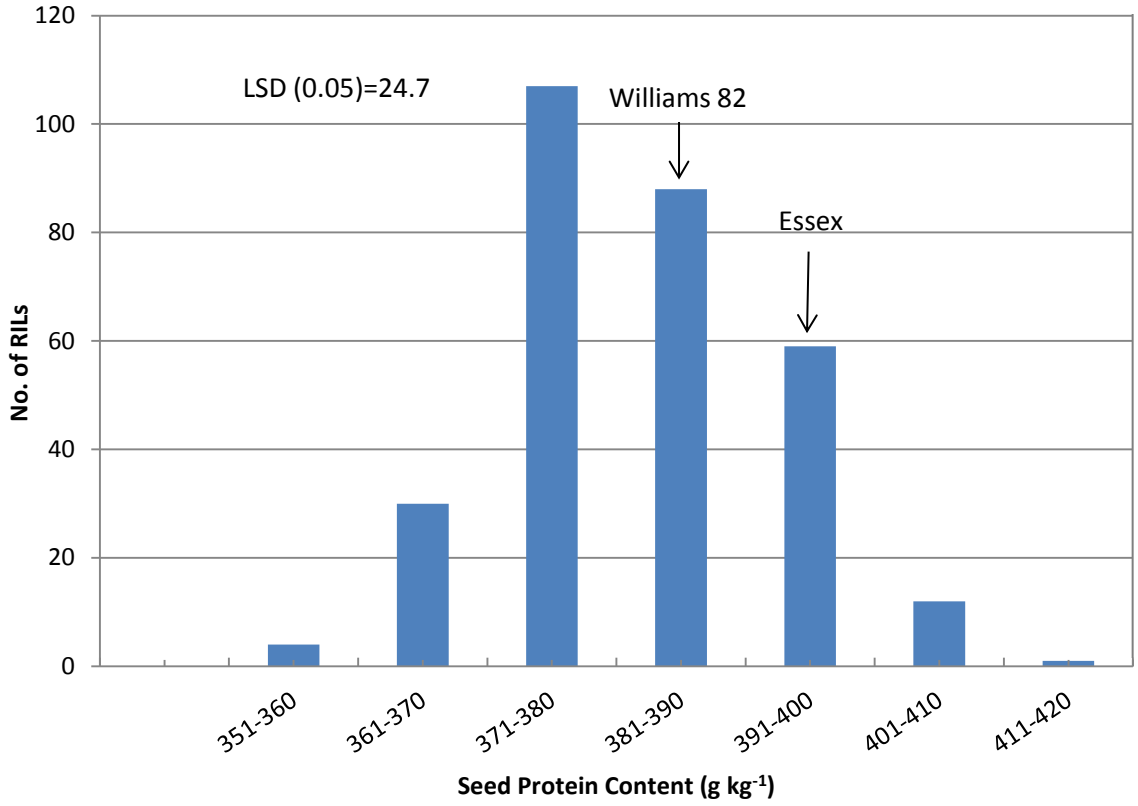

Figure 2.1 Frequency distribution of seed protein concentration for 302 $F_{5:11}$ recombinant inbred lines of Essex *x* Williams 82 averaged over three environments**.**

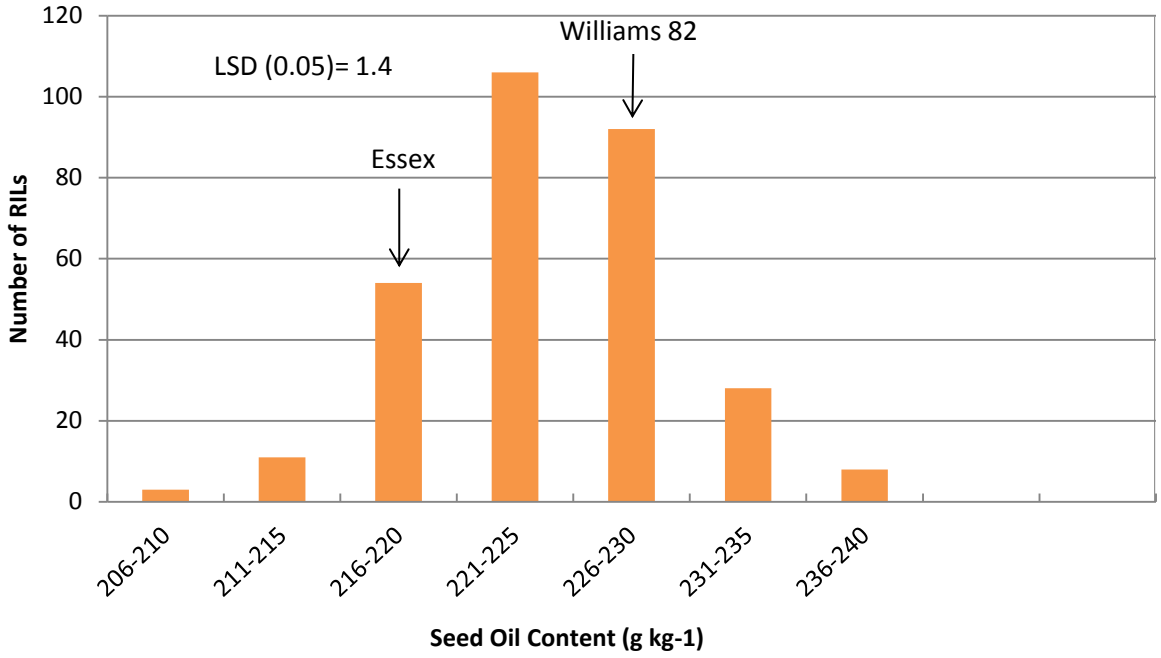

Figure 2.2 Frequency distribution of seed oil content for 302 $F_{5:11}$ recombinant inbred lines of Essex *x* Williams 82 averaged over three environments.

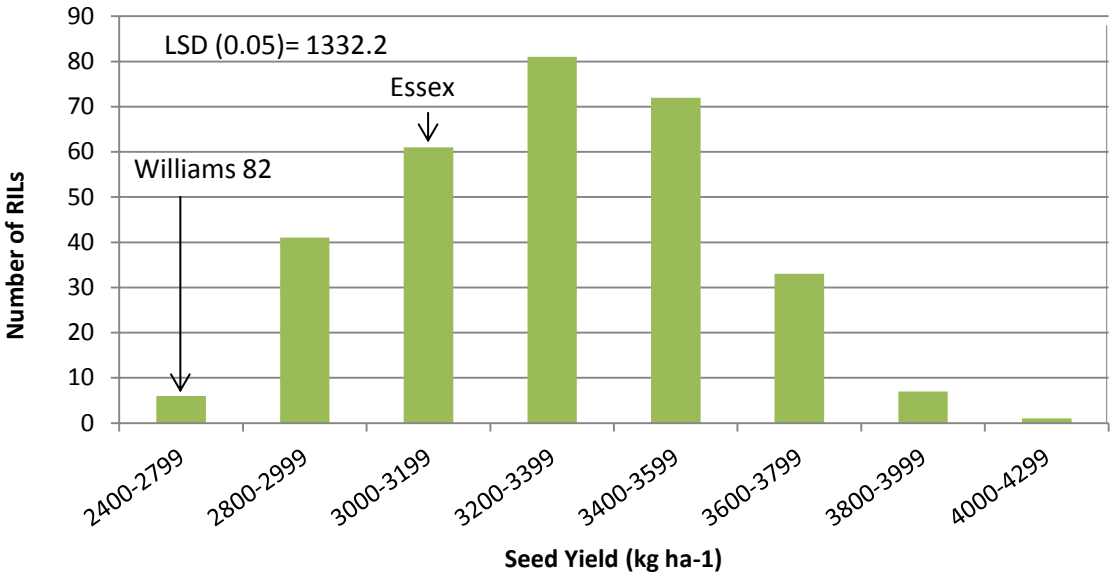

Figure 2.3 Frequency distribution of seed yield for 302 $F_{5:11}$ recombinant inbred lines of Essex *x* Williams 82 averaged over three environments.

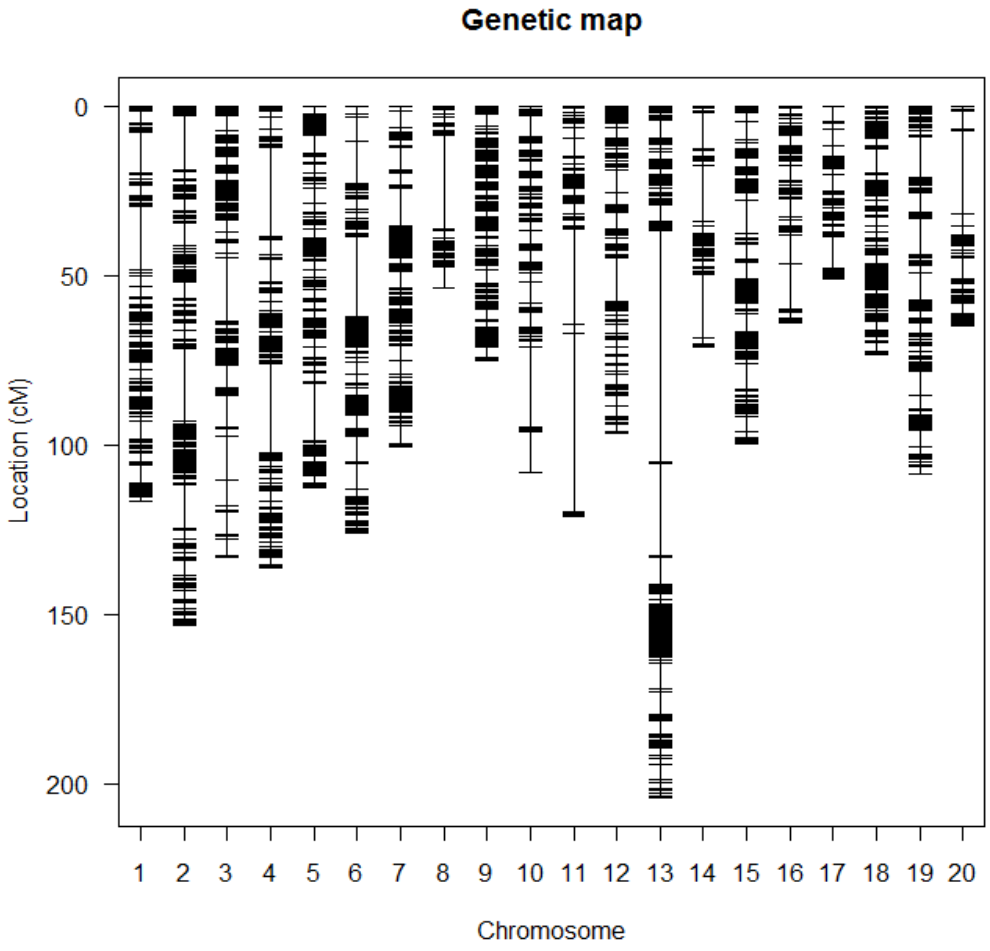

Figure 2.4 Molecular map of the Essex *x* Williams 82 $F_{5:11}$ population of 302 recombinant inbred lines mapped with 12,732 linked SNPs.

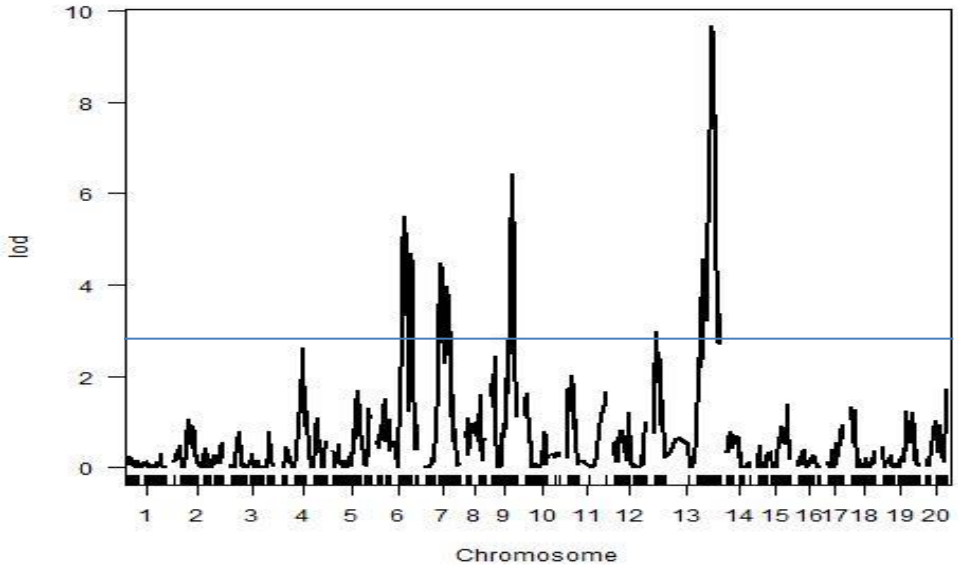

Figure 2.5 Composite interval mapping of protein for the Essex *x* Williams $F_{5:9}$ recombinant inbred lines population. Threshold indicates QTL on chromosomes 6, 7, 9, and 13. Logarithm of odds threshold =3.0

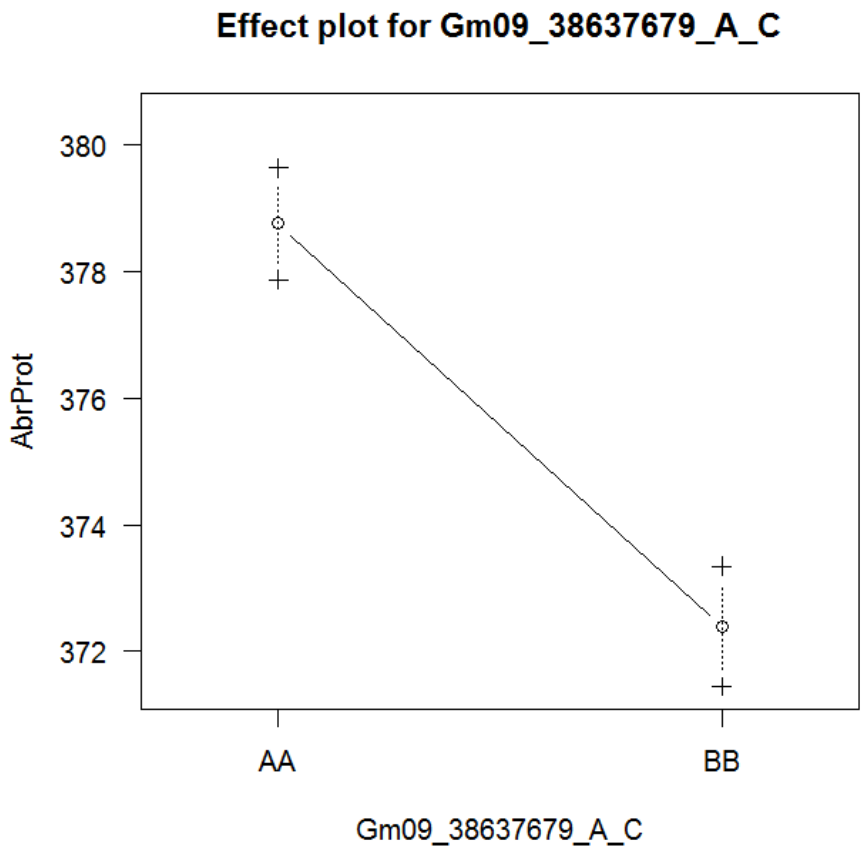

Figure 2.6 Effect plot of SNP marker Gm 9_38637679_A_C. The Williams 82 alleles are designated as AA and Essex alleles are designated BB. Figure shows the increase in protein concentration contributed by the Williams allele at this locus.

Chapter 3: Evaluation of QTL for vital amino acids for soymeal nutrition: cysteine, methionine, lysine, threonine and tryptophan in a RIL population of Essex *x* Williams 82

J. S. Abrams, B.D. Fallen, C. Nyinyi-Hatcher, A.M. Saxton, C.A Beyl, D.A. Kopsell, J.H. Orf , A.S. Killam, P. Cregan, D. Hyten and V.R. Pantalone

Abstract

Soybean is an excellent source of plant protein. The primary purpose of protein meal is to provide adequate amino acids for animal nutrition. The United States (US) is the primary producer of soybean in the world. The animal nutrition industry has called for improvement of amino acid traits in soymeal. The objectives of this study were to identify, verify and confirm quantitative trait loci (QTL) for cysteine, methionine, lysine, threonine and tryptophan and test the stability of quantitative trait loci (QTL) across multiple environments for the 'Essex *x* Williams 82' recombinant inbred line (RIL) population. For this study $F_{5:8}$ RILs were phenotyped as progeny rows in one environment and genotypic data from >50,000 SNP markers were used to identify initial QTLs. A total of 17,232 SNP markers were polymorphic. A linkage map was constructed using R/qtl and QTLs were detected using composite interval mapping. The $F_{5:8}$ and $F_{5:11}$ populations were phenotyped in for cysteine, methionine, lysine, threonine and tryptophan in one environment and multiple environments, respectively. Genotypic data were used to verify previously reported QTLs and to identify additional amino acid QTLs. Threonine and tryptophan had high heritability across multiple environments; however heritability was low to moderate for cysteine, methionine and lysine. Most amino acids were positively correlated; however seed protein and amino acids were negatively correlated. Genotype was significant for all amino acids ($p< 0.0001$) and genotype x environment interaction was significant for all amino acids except methionine. Based on testing in one environment, 15 amino acid QTLs were detected. In 2013, a total of 17 amino acid QTLs were identified. Four QTLs were verified between the two years of data, verified QTLs were found for cysteine on Gm 9 and Gm 13. The QTLs explained 4.5% to 6.8% of the variation in cysteine and higher values were conferred by Williams 82. Verified QTLs for lysine were detected on Gm 9 and Gm19. The lysine QTLs explained 3.0 to 22.4% of variation in lysine. Based on our research findings some lines have higher amino acid content and should be selected to improve amino acid composition.

KEYWORDS: amino acids, recombinant inbred lines, quantitative trait loci

Introduction

Soybean [*Glycine max* (L.) Merr.] is one of the most valuable and important agronomic crops grown in the world. Soymeal is classified as the most complete plant protein, however synthetic amino acid supplements must still be added to achieve the desired nutritional profile for animal feeds (Young and Pellett, 1994). Soymeal is primarily utilized by the chicken (poultry) (*Gallus gallus*) and swine **(***Sus scrofa domesticus*) industries. However, soymeal is also an important component in the diet of cattle (*Bos taurus*) and aquatic animals such as lobster (*Homarus americanus*) and farm- raised catfish (*Ictalurus punctatus*). In 2013, soymeal consumption in the US (26.5 million metric tons) was as follows: poultry consumed 50%, swine consumed 26%, cattle consumed 19% and 5% was consumed by aquatic animals and pets.

The primary function of protein in animal nutrition is to supply adequate amounts of the required amino acids (Friedman and Brandon, 2001). Amino acids are classified as essential and non- essential. Non- essential amino acids can be manufactured by the body, yet essential amino acids must be supplied through the diet. Five amino acids have been identified by the National Association of Animal Nutritionists (NAAN) as the most important for poultry and swine diets: cysteine, methionine, lysine, tryptophan and threonine (Boisen, 2003). Cysteine is classified as a non-essential amino acid in the poultry diet because it can be produced by the body at necessary levels. However, cysteine is an essential amino acid in the swine diet. Soymeal is currently supplemented for all the aforementioned amino acids. Certain amino acids are supplemented at different levels according to the growth stage of the animal. Lysine, threonine and tryptophan are most limiting amino acids in juvenile swine. For poultry, methionine, lysine and threonine are the most limiting. The amino acid supplementation industry has made it a primary focus to provide supplements that could eventually result in the use of less soybean meal to provide

animal nutrition. Livestock feed suppliers spend an estimated $100M in amino acid supplements annually and supplementation is not without its own problems. For example, methionine supplementation can cause bacterial degradation of volatile sulfides during soymeal processing and the sulfides can leach into the ground causing soil contamination (George and De Lumen, 1991).

As plant breeders, we have an opportunity to utilize phenotypic and molecular strategies to develop cultivars with improved amino acid composition. Improvements would decrease or eliminate the need for supplements. Such strategies would simultaneously ensure a positive impact on the soybean supply chain and aid in protecting the environment.

Amino acids are important for poultry and swine development and the soymeal industry is the primary customer of soybean farmers (Soystats, 2014). Research has been conducted in the areas of protein and oil improvement in soybean and the inverse relationship between seed protein and seed oil is established knowledge (Specht et al., 2001). However, protein concentration and amino acid composition have been found to be negatively correlated in some studies (Panthee et al., 2005, Warrington, 2011). Breeders are facing a dilemma of how to effectively improve amino acid composition without inherently decreasing protein concentration in soybean. If genotypes with high amino acid composition occur while maintaining acceptable seed protein and oil content, we will be able to produce soybean cultivars that produce soymeal that is nutritious and more profitable to the animal feed industry because fewer amino acid supplements will be required.

Few studied have been directed to elucidate the underlying genetic factors for amino acid composition in soybean. Breeding strategies such as phenotypic selection have proven to be useful in identifying superior soybean genotypes for amino acid composition. Warrington, et al.

(2014) utilized 140 F_5- derived RILs developed from a cross of 'Benning' and 'Danbaekong' to study amino acid composition. Variation was studied across five environments and genotype effects were significant ($p<0.0001$) for all traits (protein, cysteine, lysine, methionine and threonine). Genotypic mean differences of 2.5% for lysine and cysteine were detected. Tryptophan was not evaluated in the study.

Transgenic enhancement of amino acids has been attempted without success mainly due to genetic instability and regulatory controls (Altenbach et al., 1989). However, molecular strategies such as quantitative trait loci (QTL) analysis have successfully detected genomic regions controlling amino acid composition in soybean. In a molecular study of amino acid composition, a large effect QTL was detected on Gm 20 that explained 55% of variation in protein concentration, but also resulted in reduced amino acid levels (Warrington, 2011). The use of marker assisted selection was suggested to improve amino acid composition. Amino acid QTLs were also found using a population of 101 F_6 –derived recombinant inbred lines (RILs) developed from a cross between N87-984 x TN93-99 to screen 94 polymorphic SSR markers (Panthee et al. 2006a). Several amino acid QTLs were linked to previously identified microsatellite (Satt) genetic markers (Satt143, Satt168, Satt203, Satt274 and Satt495). Panthee et al. (2006b) conducted additional studies using the same population and found four QTLs controlling methionine and three QTLs controlling cysteine. Differences in methionine concentrations were found with a range of 5.1 to 7.3 g kg^{-1} seed dry weight for cysteine and 4.4 to 8.8 g kg^{-1} seed dry weight for methionine. The QTLs associated with cysteine, near Satt235, Satt252, Satt427 and Satt436 were located on *Glycine max* chromosome (Gm) 1, 13 and 18, respectively. Three QTLs associated with methionine near Satt252, Satt564, and Satt590 were on Gm 7, 13 and 18, respectively (Panthee et al., 2006b). Gm 13 and Gm 18 could have a

critical role in amino acid improvement for soybeans (Panthee et al., 2006b). In a similar study, Panthee et al. (2004) used the RIL population to map QTL that were associated with the 7S and 11S fraction of soybean storage proteins. These fractions are associated with the glycinin faction of the storage proteins and contain higher sulfur-containing amino acids. QTLs were identified on Gm 17, 19 and 20. Fallen (2012) located one QTL on Gm 13 that was associated with methionine synthesis and synthesis of eleven other amino acids. In addition, three other genomic regions on Gm 13 at positions (4.9, 21.5, 40.7 cM) were found to control multiple amino acids. A QTL associated with threonine was linked to BARC-048619 (79.1 cM) and QTL associated with methionine linked to BARC-042449 (77.4 cM) were detected on Gm 9. Both QTLs occurred within a 2 cM distance. Carson (2011) identified multiple QTL associated with protein concentration and amino acid composition. Using an F_4-derived soybean populations of NC-Roy *x* Prolina and NC-Roy x NC-106, analysis showed common regions across both populations. Genomic regions on Gm 3, 6, 8 and 9 produced QTLs for the majority of the amino acids. A QTL on Gm 10 at 88.1 – 93.2 cM had a significant effect on cysteine. These findings provide evidence that important genes regulating protein and amino acid composition may be located on Gm 9, Gm 10 and Gm 13. There are several seed amino acid QTLs listed in Soybase: six methionine, two cysteine, three lysine, five threonine and six tryptophan. There are no confirmed amino acid QTLs listed (Soybase, 2015).

The objectives of this study were: 1) Detect and verify/or confirm QTLs in Essex *x* William 82 50K RILs for amino acids: cysteine, methionine, lysine, threonine and tryptophan; 2) Determine the correlation between amino acids and seed quality traits.

Materials and Methods

Plant Materials

The University of Tennessee ExW82-50K RIL mapping population was used to conduct the current research study. The research population was derived from a cross between Essex 86-15-1 (E) and Williams 82-11-43-1 (W82); the last five numbers refer to re-selections conducted at University of Tennessee. Essex is a southern cultivar with purple flower color and gray pubescence. It has a determinant growth habit and is classified as a maturity group V soybean (Smith and Camper, 1973). 'Williams 82' is a cultivar which carries a resistance gene for *Phytopthora*; it is derived from the northern cultivar 'Williams' and has white flowers and tawny pubescence. It has an indeterminate growth habit and is classified as a maturity group III soybean (Bernard and Lindahl, 1972).

The nomenclature Essex 86-15-1 refers to a within line reselection process using the 86^{th} plant, growing the seeds from that plant as a single row then choosing the 15^{th} plant, followed by growing seeds from that plant as a single row and choosing the 1^{st} plant. The same process was used to form Williams 82-11-43-1. In 2005, the seeds were planted in the crossing block and a genetic cross of the reselected lines designated Essex 86-15-1 *x* Williams 82-11-43-1 was made at the University of Tennessee, East Tennessee Research and Education Center (ETREC) in Knoxville. The cross was designated as the ExW82-50K population. The F_1 seeds from the ExW82-50K cross were harvested and grown at the USDA Tropical Agricultural Research Station (TARS) in Isabela, PR. The F_2 population was advanced to the F_5 generation through the single seed descent method (Brim, 1966). The F_2 and F_3 generations were grown at ETREC in 2006 and 2007, respectively. The F_4 generation was grown at TARS during the spring of 2008. In the summer of 2009, $F_{4:5}$ seeds were planted and became F_5 plants in Beltsville, MD at a

USDA greenhouse. Each plant was tagged for identification and leaf tissue was collected. Seeds were harvested from 1021 plants. The $F_{5:6}$ seeds were planted in Homestead, FL in fall 2009 for a seed increase. The seeds were harvested and planted in the spring 2010 at ETREC as $F_{5:7}$ lines and seeds were harvested from each individual line. $F_{5:8}$ seeds were sent to Homestead, FL for a seed increase and $F_{5:9}$ seeds were harvested. In 2011, the $F_{5:9}$ seeds were planted for a multi-location experiment and $F_{5:10}$ seeds were harvested. Due to discrepancies in USDA-ARS flower color data and 2010 field data, flower colors were verified by growing the lines in the greenhouse at the University of Tennessee in fall 2012. After verification, the $F_{5:10}$ remnant seeds were sent to Homestead, FL for a seed increase and produced $F_{5:11}$ generation seeds. The $F_{5:11}$ seeds were planted in spring 2013 for the current research study.

Field Methods

Soybean entries from the 2010 study were planted in in Knoxville, TN at ETREC. Each line was planted as one rep in a two row plot with row length of 6 m and rows spacing of 76 cm. In 2013, the initial ExW82-50K mapping population (1021 plants) was subdivided into three categories based on maturity: Maturity Group (MG) III, MG IV and MG V. The MG V population included 302 lines that were selected to conduct this research study. Additionally three yield checks were utilized [(Osage (Chen et al., 2007), 5002T (Pantalone et al., 2004) and Ellis (Pantalone, 2015)]. Soybean entries for the 2013 study were planted in 6.1 m length two-row plots in a randomized complete block design (RCBD). Research plots were established at three locations with three replications. ETREC is located in Knoxville, TN (35.53°N 83.57°W). The soil at this location was classified as Etowah loam and the annual average rainfall is 1193.8 mm. Highland Rim Research and Education Center (HRREC) is located in Springfield, TN (36.28°N 86.51°W). The soil type at HRREC is classified as either Dickson silt loam or Sango

silt loam and the annual average rainfall is 1244.6 mm. The Research and Education Center at Milan (RECM) is located in Milan, TN (35.54° N 88.44° W). The soil is classified as Loring B2 series fine silt and annual average rainfall is 1371.6 mm.

Agronomic Trait Evaluation

In 2010 and 2013 all experimental plots were evaluated for agronomic traits and seeds were monitored for germination rates. Phenotypic data were collected for flower color, pubescence color, lodging, height and maturity. Flower color was noted as purple, white or segregating when 95% of the plants had bloomed. Plant height and lodging were measured at maturity. Lodging was scored on a 1-5 scale; with 1 representing plants that were upright and 5 represented plants that were prostrate. Maturity was recorded when 95% of the pods had achieved their mature color. Pubescence was scored as gray, tawny or segregating, when 95% of the pods in a plot showed their mature color (Fehr and Caviness, 1977). All plots were rogued to ensure genetic integrity.

For the 2010 study, maturity dates were recorded according to the Julian calendar with day 1 being January 1st. The first maturity was recorded on day 251 (Sept 8, 2010) and the last maturity was recorded on day 288 (Oct 15, 2010). The RIL lines were placed into MG III through MG V based on maturity dates. In 2013, maturity was recorded with Day 1 as September 1. Because all lines were classified as MG V, the first maturity date was recorded on Sept 27th, corresponding to day 27 in the maturity log. These 2010 maturity dates were converted to a Sept 1 start date for the research data analysis.

Near Infrared Analysis for Amino Acid Composition

Phenotypic data for amino acid composition were collected for each research plot. Whole bean subsamples weighing 25 g were ground using a water-cooled Knifetec 1095 Sample Mill (FOSS Tecator, S-26321, Hogana, Sweden). The samples were ground for 20 seconds and placed into Whirl Pak bags (Nasco, Fort Atkinson, WI). All samples were barcode labeled using the ZM 1000 Barcode Printer System (Zebra Technologies, Lincolnshire, IL). The ground samples were analyzed at the Soybean Analysis Lab at the University of Minnesota, St. Paul, MN (Dr. Jim Orf). The laboratory was equipped with a JET Air Filtration System (LaVergne, TN) to minimize particulate matter from the ground soybean samples. A subsample weighing 12.5 g was placed into a small sample cup to conduct the ground bean analysis. The sample was leveled using a spatula and placed into the Perten near infrared (NIR) analyzer (Hägersten, Sweden). The Perten software conducted an initial scan of auto diagnostics for instrument response, wavelength accuracy and NIR repeatability. The scan reported the soybean composition percentages. In prior publications amino acid content was measured by the percentage of total seed content. However, Warrington (2011) described a more accurate method to access amino acid composition, describing it as a portion of total crude protein. This method had been adopted by the soybean breeding community as a more accurate calculation of actual amino acid composition. The NIR percentages were converted to obtain predicted cysteine (Cys), lysine (Lys), methionine (Met), threonine (Thr) and tryptophan (Trp) in g kg^{-1} of crude protein.

Genotyping

F_5 plants were tagged for identification and DNA was extracted from each F_5 greenhouse plant grown at the Soybean Genomics Laboratory at the USDA Beltsville Agricultural Research

Center (USDA-ARS) in Beltsville, MD. Samples contained 50 µl [microliters] of DNA at a 200 ng/µl content. The samples were assayed using the GoldenGate® assay with >50,000 SNP markers following the protocol from the manufacturer and methods described by Hyten et al. (2008) and Fan et al. (2003). All samples were assayed using the Illumina BeadStation 500G (Illumina, San Diego, CA). The population produced 17,232 polymorphic SNP markers. Data from the Beltsville analysis were used for the genetic analysis component of the current research population.

Data Analysis

Analyses of variance (ANOVA) were conducted on the phenotypic data to determine if there were significant differences among the RIL genotypes, environment and genetic x environment (g x e) interaction. All ANOVA were conducted using the MIXED procedure of SAS (SAS ver. 9.3, Cary, NC). ANOVA were conducted to detect significant differences among RILs for amino acid composition. Initially, all effects were tested as random factors and included genotype, environment, genotype x environment and genotype x replication within environment. To allow greater precision and higher power, replication within environment was removed from the model. The initial model was found to be too conservative and negated a high proportion of measurable variation. The final model was the following:

$$Y_{ij} = \mu + B_i + T_j * + B*T_{ij}$$

where, Y is the observation of the *j*th treatment (genotype) in the *i*th block (environment).
A second analysis was run using genotype as a fixed term and environment and genotype x environment as random terms. Fisher's LSD was performed for mean separation with the MIXED procedure of SAS (SAS ver. 9.3, Cary, NC). The least squares means were calculated and compiled to use in QTL analysis.

Pearson's correlation analysis was performed to determine correlations among seed protein, seed oil and yield. Phenotypic correlations were determined using CORR procedure (SAS ver. 9.3, Cary, NC). To determine the portion of phenotypic variation among RILs that resulted from genetic differences and estimate heritability, restricted maximum likelihood estimation was used. A broad sense estimate of heritability of cysteine, methionine, lysine, threonine and tryptophan in the population was calculated on an entry mean basis (Nyquist, 1991) as follows:

$$h^2 = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_{ge}/e + \sigma^2/re}$$

where, h^2 represents the heritability, σ^2_g is genotypic variance, σ^2_{ge} is genotype x environment variance, σ^2 is error variance, *r* is number of replications and *e* is number of environments. Since the population was F_5-derived, most of the genetic variance was additive. Therefore, we are obtaining an approximation of narrow sense heritability with this formula.

Linkage Map and Quantitative Trait Loci Analysis

R/qtl (Broman and Sen, 2009) was used to construct a genetic linkage map using 17,232 SNP markers and 302 genotypes. The estimated map length was 2072 cM and utilized 12,730 markers, after unlinked markers were discarded. Chromosomal location, marker order and position were determined by composite interval mapping (CIM) (Broman, 2001; Broman and Sen, 2009). A standard walking speed of 2 cM was used to conduct CIM (Broman and Sen, 2009). Ten thousand permutations were performed to establish a log odds (LOD) threshold of 3.0 at α=0.001 (Müller-Myhsok, 2009). The LOD threshold of 3.0 was consistently applied to identify amino acid QTLs on each of the twenty soybean chromosomes.

Results and Discussion

Phenotypic Analysis of Amino Acid Composition

Multi- environment testing is useful to establish the genetic value of a genotype. Based on testing in three environments, significant differences were observed among RIL genotypes for amino acid composition for cysteine, methionine, lysine, threonine and tryptophan ($p<0.0001$). The differences among environments were not significant for cysteine ($p=0.1597$), methionine ($p=0.1601$), lysine ($p=0.1699$), threonine ($p=0.1594$) or tryptophan ($p=0.1593$). However, genotype x environment interactions was significant for all amino acids ($p<0.0001$) except methionine ($p=0.0789$). As parents, Essex and Williams 82 are similar in amino acid composition, differing < 1.0 g kg^{-1} in crude protein for all amino acids tested (Table 3.1). Higher amino acid composition was observed in Williams 82 for all amino acids except threonine, which was higher in Essex.

Phenotypic correlations were observed among all RIL genotypes ($p<0.01$). Most amino acids were positively correlated with one another (Table 3.2). Only cysteine and tryptophan were negatively correlated ($r=-0.01$), but the amount was negligible. Cysteine and methionine had a moderate positive correlation ($r=0.66$), however, cysteine had a weak positive correlation with lysine ($r=0.17$) and threonine ($r=0.31$). A strong positive correlation was observed between lysine and threonine ($r=0.78$). Methionine had a positive moderate correlation with lysine ($r=0.51$) and threonine ($r=0.61$), but methionine had a weak positive correlation with tryptophan ($r=0.33$). Lysine had moderate to strong correlations with tryptophan ($r=0.62$) and threonine ($r=0.78$). Threonine and tryptophan, however had a positive moderate correlation ($r=0.48$).

Amino acids are a vital component of seed protein and they are affected by seed protein and seed oil ratios within the seed (Wilson, 2004). Therefore, seed protein and seed oil content

were monitored with respect to amino acid composition. Seed protein was negatively correlated with each amino acid tested. There was a strong negative correlation between seed protein and lysine (r=-0.80) and seed protein and threonine (r=-0.88). Moderate correlations between seed protein and methionine (r=-0.57) and seed protein and tryptophan (r=-0.64) was observed. There was a weak negative correlation between seed protein and cysteine (r=-0.20). Seed oil had a weak to moderate positive correlation with all five of the amino acids tested (r=0.08 to 0.56).

All data for amino acids are described in grams of amino acid per kilogram of crude protein (cp) within the seed. Cysteine ranged from 11.0 g kg^{-1}cp to 21.3 g kg^{-1} cp (Table 3.3). The maximum differed from the mean by 10.3 g kg^{-1} cp. For methionine, RILs ranged from 12.9 g kg^{-1} cp to 18.1 g kg^{-1} cp. Lysine composition ranged from 63.2 g kg^{-1} cp to 73.7 g kg^{-1} cp, a difference of more than 10.0 g kg^{-1} cp. Smaller differences were observed in threonine and tryptophan. Threonine ranged between 37.5 g kg^{-1} cp to 43.9 g kg^{-1} cp. RILs ranged between 10.2 g kg^{-1}cp to 12.7 g kg^{-1} cp for tryptophan. Heritability estimates were on the extremes of high and low values for most amino acids. Cysteine (29.8%) and methionine (53.9%) heritabilities were lower and lysine (59.3%) heritability was higher than values previously observed in the Essex x Williams 82 RIL population (Fallen, 2013). Threonine and tryptophan had high heritability values (78.0% and 81.6%, respectively) (Table 3.3). Values did not follow the trend of amino acid heritability observed in previous studies with different parentage (Panthee, 2006a; Carson, 2011). Heritability can differ between lines due to genetic affects. Heritability is important because it is often utilized to determine the merit of a particular breeding strategy.

All amino acids included in this study showed significant differences for genotype. Genetic x environment interaction was detected for all amino acids except methionine, which

may indicate a stronger genetic factor underlying the expression of methionine. The results show a positive correlation among most of the amino acids. There was a negative correlation between seed protein concentration and all of the amino acids associated with this study. Similar finding were reported by Carson (2011), who concluded that it would be unlikely to be able to improve amino acid composition and seed protein simultaneously.

The sulfur-containing amino acids (methionine and cysteine) are highly supplemented in in soymeal to provide adequate nutrition for animal diets (Allee, 2005). Lysine, threonine and tryptophan supplements are adjusted to different levels depending on growth stages of poultry and swine (Boisen, 2003). The United Soybean Board (USB) has put forth several goals regarding amino acid improvement in soybean, asking breeders to make 1.1-1.23 X improvements in soybean meal traits for the five amino acids included in this study. Transgressive segregates with an increase in amino acid composition were identified among the RILs in this study (Figures 3.1, 3.2, 3.3, 3.4, 3.5). Some of the genotypes produced over 5.2 g kg^{-1} cp. -10.3 g kg^{-1} cp for a given amino acid. Breeding cultivars with improved amino acid composition will require attentive observation, but can be accomplished by using a combination of phenotypic and molecular breeding strategies.

Quantitative Trait Loci Analysis of Five Vital Amino Acids

Initial analysis of marker data was conducted using R/qtl (Broman and Sen, 2009). A total of 12,732 linked markers were included in the map which covered a distance of 2032 cM and included 20 linkage groups. Composite interval mapping was conducted to detect all QTLs. Ten thousand permutations were used to establish a LOD threshold of 3.0 and a type I error rate of at least $p< 0.01$. For initial QTL discovery, data were analyzed for Essex x Williams 82 RILs grown in 2010 in Knoxville, TN. Analysis revealed QTLs for cysteine on Gm 3, 9 and 13. The

individual QTL explained 5.4% to 7.6 % of variation in cysteine. Alleles for increased cysteine levels were contributed by both Essex and Williams 82 at different chromosomes (Table 3.4). The QTL effect is reported in terms of homozygous allelic pairs of Williams 82 (W) or Essex (E) at a locus, because the population is F_5 derived. For example, RILs homozygous for the Essex allele demonstrated an increase of 0.20 g kg^{-1} of cysteine at the locus on Gm 3 (Table 3.4). Four lysine QTLs were identified in Gm 5, 9, 13 and 19. The QTL on Gm 19 located at 201.0 cM was highly significant (R^2= 14.2) and this major QTL that explained 22.4% of the variation in lysine. There were two QTL detected for methionine. The QTLs were located on Gm 9 (48.0 cM) and Gm 19 (202.4 cM) and the individual QTL explained 5.8% and 10.8% of variation in methionine, respectively. Four threonine QTLs were detected on Gm 13, 14, 19 and 20. The QTL on Gm 19 had a large peak (R^2=23.6) and the QTL explained 21.1% of the variation in threonine. Most of the alleles to increase threonine were from Williams 82. Two tryptophan QTLs were also identified on Gm 9 (47.0 cM) and Gm 19 (31.2 cM). Results for 2010 provided further justification for the expanded study in 2013, which was conducted in multiple environments.

In 2013, QTL analysis was conducted using two methods to evaluate the Essex x Williams 82 $F_{5:11}$ RIL population. All location and individual location QTL analyses were conducted. The all location analysis combined phenotypic data from each of the three individual environments, while the individual locations analysis kept phenotypic data from each environment separate which allowed evaluation of QTL stability across various environments and observation of the environmental impact on the presence of a particular QTL.
An all location analysis is the most typical analysis conducted for QTL in multiple environments and is the primary focus of the QTL analysis in this study.

Cysteine QTLs were identified on Gm 9 (0.0 cM) and Gm 13 (199.0 cM), the individual QTL explained 5.8% and 4.5%, of the variation, respectively (Table 3.5). Both QTLs were conferred by the Williams 82 allele. Five lysine QTLs were identified on Gm 6 (57.2 cM), 7 (48.7 cM), 9 (60.0 cM), 13 (184.0 cM) and 19 (196.0 cM). Individual QTL explained 5.4% - 12.0% of the variation in lysine (Table 3.5). Three methionine QTLs were identified on Gm 9 (0.0 cM), Gm 13 (184.0 cM) and Gm 18 (5.1 cM). All QTLs observed for methionine were conferred by the Williams 82 allele. The individual QTL explained 4.0% - 7.4% of the variation in methionine. Four threonine QTLs were identified on Gm 6 (58.0 cM), 7 (43.3 cM), 9 (64.1 cM) and 13 (183.0 cM). They explained 5.7% - 14.3% of the variation in threonine and the QTL on Gm 13 was a major QTL. Three tryptophan QTLs were observed on Gm 6 (54.1 cM), 13 (188.3 cM) and 19 (201.4 cM). They explained 3.8% - 7.3% of the variation in tryptophan. QTLs observed for tryptophan were conferred by both the Essex and Williams 82 alleles. Verified QTLs in this study are those which: a) have been detected in both years, b) are located on the same chromosome, c) are within 10 cM apart in distance and, d) have been conferred by the same allele in 2010 and 2013 studies. A verified QTL for cysteine was detected on Gm 9 at 0.0 cM in 2010 at 9.0 cM in 2013. An additional verified QTL was found for cysteine on Gm 13 in 2010 at 204.3 cM and at 199.0 cM in 2013 (Table 3.4, Table 3.5). The QTL explained 4.5% to 7.6% of the variation in cysteine and were conferred by Williams 82. Two lysine QTLs were verified QTL on Gm 9 at 60.0 cM in 2010 and 64.1 cM in 2013, Gm19 at 201.1 cm in 2010 and 196.0 cM in 2013. However, there was a large difference in the amount of variation explained by the QTL on Gm 19. The 2010 QTL explained 22.4% of variation in lysine, the 2013 QTL explained 5.4% of the variation in lysine. Overall, 15 amino acid QTLs were identified in 2010

and 17 amino acid QTLs were identified in 2013. A total of 4 amino acid QTLs were verified in this study.

Initial reporting of data from the individual location analysis was for verified QTLs only. A total of 41 multi-environment QTL were identified in 2013 (Table 3.6). A cysteine QTL was verified on Gm 13, it was detected at 204.3 cM in 2010 and at 199.0 cM in 2013. The QTL was not present in any other environment. A lysine QTL was verified on Gm 9 in two of the three environments at 58.0 cM to 61.0 cM. The individual QTL accounted for 3.8% to 5.8% of variation in lysine. Additional QTLs were verified on Gm 13 in two of the three environments at 186.6 cM and 185.0 cM. The individual QTL accounted for 4.8% to 7.9% of variation in lysine. A QTL was verified on Gm 19 at 194.0 cM and the individual QTL explained 3.3% of the variation in lysine. No verified QTLs for methionine were detected across multiple years, however a methionine QTL on Gm 9 appeared in two environments and the two QTLs were documented less than 5 cM apart in distance. The individual QTL values explained 3.5% to 4.9% of variation in methionine. Threonine QTLs on Gm 13 were verified in both years and across all environments in 2013. The QTLs were conferred by Williams 82 and the individual QTL explained 5.5% to 7.6% of variation in threonine. Additionally, two QTLs were detected in two environments on Gm 6 (55.0 cM and 58.0 cM) and Gm 19 (196.6 cM and 196.7 cM).

The combined QTL analyses produced verified QTL for cysteine and lysine. No QTL were verified for methionine, threonine or tryptophan in the all location analysis. A cysteine QTL on Gm 9 was verified in at least one location in all analyses. This is the type of QTL we hoped to identify because of its presence in multi-year, combined and individual QTL analysis (Tables 3.4, 3.5 and 3.6). In all cases, the QTL explained 6.3% to 6.8% of the variation in cysteine. However, the absence of the QTL in some environments may indicate the g x e

influence of the expression of this QTL. The lysine QTL on Gm 9, present in both years and two environments, explained 3.0 % to 8.5 % of variation in lysine (Tables 3.4, 3.5 and 3.6). The ranges reported are also evidence of the varied effects that the same QTL can have on expression of a particular amino acid. There is evidence that the QTL located on Gm 13 has pleiotropic effects on protein and several amino acids. There were no positionally confirmed QTL for amino acids.

Conclusions

The individual location analysis was beneficial for detecting QTLs that appeared in some environments and were absent in others. The results show inconsistency of some QTLs across environments. Breeders will have to continue to rely on phenotypic information and incorporate marker assisted selection for those amino acid QTLs that have been verified for improving amino acid composition. Breeders should also consider epigenetic analysis of lines that fail to express or have higher expression of some amino acids. Such analysis could produce helpful information to aid in amino acid improvement. In addition, breeders need to understand the biochemistry of amino acid synthesis to build better genetic populations for confirmation of genomic regions governing key amino acids. The most limiting amino acid (methionine) is one of the final amino acids synthesized at the end of the aspartate pathway (Figure 3.6). Knowledge of genetic control leading up to its synthesis may greatly improve our overall knowledge of soybean amino acids expression and provide efficiency in making further improvements in soybean amino acid composition.

References

Allee, G. 2005. Protein and Amino Acid Sources for Swine Diets. National Swine Nutrition Guide:30-36.

Altenbach, S.B., K.W Pearson, G. Meeker, L.C. Staraci and S.S.M. Sun. 1989. Enhancement of the methionine content of seed protein s by the expression of a chimeric gene encoding a methionine-rich protein in transgenic plants. Plant Mol Biol 13:513-522.

Bernard, R.L. and C.R. Cremeens. 1988. Registration of 'Williams 82' Soybean. Crop Sci 28:1027-1028.

Bernard, R.L. and D.A. Lindahl. 1972. Registration of Williams soybean. Crop Sci. 12:716.

Bernardo, R. 2010. Breeding for quantitative traits in plants. 2nd Edition. p.91.Stemma Press, Woodmury, MN.

Boisen, S. 2003. Ideal dietary amino acid profiles for pigs. In: JPF D'Mello (2nd edn) Amino Acids in Animal Nutrition. Cambridge, MA, pp. 157-168.

Brim, C.A. 1966. A modified pedigree method of selection in soybeans. Crop Sci. 6:220.

Broman, K.W. and S. Sen. 2009. A guide to QTL mapping with R/qtl. Springer, New York.

Carson, C. 2011. Genetic control of protein and amino acid content in soybean determined in two genetically connected populations. Doctor of Philosophy Dissertation. North Carolina State University.

Chen, P., C. H. Sneller, L.A. Mozzoni and J.C. Rupe. 2007. Registration of Osage soybean. J. Plant Reg 1:89-92.

Fallen, B.D., C.N. Hatcher, F. L. Allen, D. A. Kopsell, A. A. Saxton, P. Chen, S. Kantartzi, P. B. Cregan, D. L. Hyten and V. R. Pantalone. 2013. Soybean seed amino acid content QTL detected using the universal soy linkage panel 1.0 with 1536 SNPs. J. of Plant Gen. Sci. 1:68-79.

Fallen, B.D. 2012. Detection of Soybean Amino Acid QTLs and Seed Yield QTLs Using Selective Genotyping Doctor of Philosophy Dissertation. University of Tennessee.

Fan, J.B., A. Oliphant, R. Shen, B.G. Kermani, F. Garcia, K.L. Gunderson, M. Hansen, F. Steemers, S. L. Butler, P. Deloukas, L. Galver, S. Hunt, C. McBride, M. Bibikova, T. Rubano, J. Chen, E. Wickham, D. Doucet, W. Chang, D. Campbell, B. Zhang, S. Kruglyak, D. Bentley, J. Haas, P. Rigault, L. Zhou, J. Stuelpnagel, M.S. Chee. 2003. Highly parallel SNP genotyping. Cold Spring Harbor Symp. Quant. Biol. 68:69–78.

Friedman M, and Brandon D. 2001. Nutritional and Health Benefits of Soy Proteins. J. Agric. Food Chem 49:1069-1086.

George, A.A. and B. O. De Lumen. 1991. A novel methionine-rich protein in soybean; Identification, amino composition and N-terminus sequence. J. Agric Food Chem 38:224-227.

Hildebrand, D. 2010. Lecture 24: Nitrogen Metabolism – Essential and non-protein amino acid biosynthesis. BCH/PPA/PLS 609 Plant Biochemistry. (http://www.uky.edu/~dhild/ biochem/24/lect24.html. Verified 2 Mar 2015)

Hyten, D.L., Q. Song, Y. Zhu, I.Y. Choi, M.S. Yoon, J.E. Specht, L.K. Matukumalli, R. L. Nelson, R.C. Shoemaker, N.D. Young, P.B. Cregan. 2008. High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor. Appl. Genet. 116: 945-952.

Müller-Myhsok. 2009. Estimating experiment-wise error rate by permutations. mga.bionet.nsc.ru/~yurii/courses/esp29.../permlecture_2009_6spp.pdf 9 (accessed. Feb 17, 2015).

Nyquist W.E. 1991. Estimation of heritability and prediction of selection response in plant populations. Crit. Rev. Plant Sci. 10: 235–322.

Pantalone, V., F. Allen, and D. Ellis. Registration of '5002T' soybean. 2004. Crop Sci 44: 1483-1484.

Pantalone, V. 2015. Registration of 'Ellis' soybean. J. Plant Registrations (in press)

Panthee, D.R., P. Kwanyuen, C.E. Sams, D.R. West, A.M. Saxton, and V.R. Pantalone. 2004. Quantitative trait loci for β-conglycinin (7S) and glycinin (11S) fractions of soybean storage protein. J of Amer Oil Chem. 88 1005-1012.

Panthee, D.R., V.R. Pantalone, D.R. West, A.M. Saxton, and C.E. Sams. 2005. Quantitative trait loci for seed protein and oil content, and seed size in soybean. Crop Sci. 45: 2015-2022.

Panthee, D.R., V.R. Pantalone, A.M. Saxton, D.R. West, C.E. Sams. 2006a.Genomic regions associated with amino acid composition in soybean. Crop Sci 17:79-89.

Panthee, D.R., V.R. Pantalone, A.M. Saxton, D.R. West, C.E. Sams, J.H. Orf, A.S. Killam. 2006b. Quantitative trait loci controlling sulfur containing amino acids, methionine and cysteine, in soybean seeds. Theor. Appl. Gen. 112:546-553.

SAS. Institute. 2013. SAS/STAT. Release 9.3 SAS Inst., Cary, NC.

Smith, T.J., and H.M. Camper. 1973. Registration of Essex soybean.

Song, Q.J., L.F. Marek, R.C. Shoemaker, K.G. Lark, V.C. Concibido, X. Delannay, J.E. Specht, and P.B. Cregan. 2004. A new integrated genetic linkage map of the soybean. Theor. Appl. Genet. 109:122–128.

Soybase and Soybean Breeder's Toolbox. http://www.soybase.org/ (accessed 28 Feb. 2015)

Soy Stats 2014. A reference guide to important soybean facts and figures. Available at http://www.soystats.com/ (accessed 3 Mar. 2015).

Specht, J.E., M. Germann, J.P. Markwell, K.G. Lark, J.H. Orf, M. Macrander, K. Chase, J. Chung, and J.L. Graef. 2001. Soybean response to water: A QTL analysis of drought tolerance. Crop Sci. 41:493–509.

Warrington, C. 2011. QTL mapping and optimum resource allocation for enhancing amino acid content in soybean. Doctor of Philosophy Dissertation. University of Georgia.

Warrington, C. V., H. A. Haleem, J. H. Orf, A. A. Killam, B. Bajjalieh, Z. Li and R. Boerma. 2014. Resource allocation for selection of seed protein and amino acids in soybean. Crop Sci. 54: 963-970.

Wilson, R.F. 2004. Seed composition. 2004. In: Boerma, H.R. and Specht, J.E. (Eds.), Soybeans: Improvement, Production, and Uses, Third Edition. ASA, CSSA, SSSA, Madison, WI, USA, pp. 621–677.

Young, V.R. and P.L. Pellett. 1994. Plant protein relation to human and amino acid nutrition. Am. Jour. Clin. Nutr. 59: 1203S-1212S.

Appendix 3: Chapter 3 Tables and Figures

Table 3.1 Mean amino acid compositions (g kg^{-1} of crude protein) and protein and oil content (g kg^{-1} of seed) of the parents and checks grown in Knoxville, TN; Springfield, TN; and Milan, TN in 2013.

	Parent Means		Check Means		
Traits	Essex	Williams 82	Ellis	5002T	Osage
Protein (DM) (g kg^{-1}) †	393.8	386.5	366.8	363.4	394.2
Oil (DM) (g kg^{-1}) †	217.1	226.2	213.8	225.4	209.8
Cysteine (g kg^{-1})‡	18.2	18.8	18.2	18.6	17.9
Methionine (g kg^{-1}) ‡	16.4	16.8	16.8	17.1	16.4
Lysine (g kg^{-1}) ‡	70.3	70.6	71.9	72.0	70.5
Threonine (g kg^{-1}) ‡	42.4	41.7	42.4	42.5	41.3
Tryptophan (g kg^{-1}) ‡	11.5	11.6	11.7	11.6	11.6

† Grams per kg of crude protein
‡ Grams per kg of seed dry matter (DM)

Table 3.2 Pearson's Correlation coefficients between major soybean seed traits and amino acids in 302 $F_{5:11}$-derived recombinant inbred lines of Essex 86-15-1 x Williams 82-11-43-1 grown in Knoxville, TN; Springfield, TN; and Milan, TN.

Trait	**Protein (g kg^{-1})** †	**Oil (gkg^{-1})** †	**Cys (g kg^{-1})** ‡	**Met (g kg^{-1})** ‡	**Lys (g kg^{-1})** ‡	**Thr (g kg^{-1})** ‡
Oil (gkg^{-1}) †	-0.70					
Cysteine (g kg^{-1}) ‡	-0.20	0.08				
Methionine (g kg^{-1}) ‡	-0.57	0.31	0.66			
Lysine (g kg^{-1}) ‡	-0.80	0.44	0.17	0.51		
Threonine (g kg^{-1}) ‡	-0.88	0.59	0.31	0.61	0.78	
Tryptophan (g kg^{-1}) ‡	-0.64	0.56	-0.01	0.33	0.62	0.48

All values were significant at $p < 0.0001$
All tests were conducted on dry weight basis
† Grams per kg of seed
‡ Grams per kg of crude protein

Table 3.3 Mean amino acid compositions (g kg^{-1} of crude protein) and protein and oil concentration g kg^{-1} of seeds from 302 $F_{5:11}$-derived recombinant inbred lines of Essex 86-15-1 *x* Williams 82-11-43-1 grown in Knoxville, TN; Springfield, TN; and Milan, TN in 2013.

Trait	Min	Mean	Max	$LSD_{0.05}$	h^2 (%)
Protein (DM) †	331.6	382.2	461.2	24.1	87.4
Oil (DM)†	193.1	224	248.5	13.3	87.2
Cysteine ‡	11.5	18.8	21.3	1.7	29.8
Methionine ‡	12.9	17.0	18.1	1.0	53.9
Lysine ‡	63.2	70.9	73.7	1.8	59.3
Threonine ‡	37.5	42.1	43.9	1.1	78.0
Tryptophan ‡	10.2	11.7	12.7	0.5	81.6

All tests were conducted on dry matter basis (DM)
* p=0.01
† Grams per kg of seed
‡Grams per kg of crude protein

Table 3.4 Quantitative trait loci identified using R/qtl for composite interval mapping located on various chromosomes associated with amino acid composition in 302 $F_{5:8}$-derived recombinant inbred lines of Essex 86-15-1 x Williams 82-11-43-1 grown 2010 in Knoxville, TN.

Location	Trait	QTL Name	Chr†	MLG‡	Molecular Marker	Loc § (cM)	LOD ¶	Confidence Interval of QTL position	R^2 (%)	Effect # g kg^{-1}
Knoxville, TN	Cysteine	Seed Cys 2-1	Gm 3	N	Gm03_1077329_C_T	9.0	4.1	6.0-14.6	6.5	0.20 (E)
Knoxville, TN	Cysteine	Seed Cys 2-2	Gm 9	K	Gm09_1879918_G_A	0.7	3.4	0.0-7.0	5.4	0.20 (W)
Knoxville, TN	Cysteine	Seed Cys 2-3	Gm 13	F	Gm13_39237614_T_C	204.3	3.8	200.3-204.6	7.6	0.20 (W)
Knoxville, TN	Lysine	Seed Lys 2-1	Gm 5	A1	Gm05_37376934_G_A	78.0	3.1	61.0-112.7	4.7	0.30 (E)
Knoxville, TN	Lysine	Seed Lys 2-2	Gm 9	K	Gm09_38956335_G_A	64.1	3.3	42.15-69.0	3.0	0.30(E)
Knoxville, TN	Lysine	Seed Lys 2-3	Gm 13	F	Gm13_1395656_T_C	9.0	4.0	4.1-201.9	6.7	0.40 (E)
Knoxville, TN	Lysine	Seed Lys 2-4	Gm 19	L	Gm19_44938831_A_G	201.0	14.2	200.3-201.6	22.4	0.70(W)
Knoxville, TN	Methionine	Seed Met 2-1	Gm 9	K	Gm09_35211690_G_A	48.0	3.8	40.0-50.0	5.8	0.20 (E)
Knoxville, TN	Methionine	Seed Met 2-2	Gm 19	L	Gm19_45369424_T_C	202.4	7.4	194.1-205.0	10.8	0.20 (W)
Knoxville, TN	Threonine	Seed Thr 2-1	Gm 13	F	Gm13_38249824_T_C	197.0	5.6	195.0-204.6	7.5	0.30 (W)
Knoxville, TN	Threonine	Seed Thr 2-2	Gm 14	B2	Gm14_30024382_T_C	45.2	3.1	35.4-54.0	6	0.30 (E)
Knoxville, TN	Threonine	Seed Thr 2-3	Gm 19	L	Gm19_45099890_G_A	201.4	23.6	200.3-204.0	21.1	0.60 (W)
Knoxville, TN	Threonine	Seed Thr 2-4	Gm 20	I	Gm20_46108934_G_A	31.4	3.6	29.2-32.9	3	0.20 (W)
Knoxville, TN	Tryptophan	Seed Tryp 2-1	Gm 9	K	Gm09_35211690_G_A	47.0	4.0	41.1- 52.0	4.3	0.09 (E)
Knoxville, TN	Tryptophan	Seed Tryp 2-2	Gm 19	L	Gm19_45369424_T_C	31.2	7.2	0.0-32.9	12.8	1.78 (W)

†Chr., chromosome.
MLG=Molecular Linkage Group
§ The QTL position was determined
¶ LOD, logarithm of the odds.
# Effect indicates quantitative change in amino acid composition associated with either (E) Essex 15-86-1 or (W) Williams 82-11-43-1 allele.

Table 3.5 Combined location analysis of quantitative trait loci identified using R/qtl for composite interval mapping located on various chromosomes associated with amino acid composition in 302 $F_{5:11}$-derived recombinant inbred lines of Essex 86-15-1 x Williams 82-11-43-1 grown in 2013 in Knoxville, TN; Springfield, TN; and Milan, TN.

Trait	QTL Name	Chr†	MLG‡	Molecular Marker	Loc § (cM)	LOD ¶	Confidence Interval of QTL position	R^2 (%)	Effect # g kg^{-1}
Cysteine	Seed Cys 2-4	Gm 9	K	Gm09_1723633_G_A	0.0	5.8	0.0-4.0	6.8	0.08(W)
Cysteine	Seed Cys 2-5	Gm 13	F	Gm13_38249824_T_C	199.0	4.4	196.0-201.9	4.5	0.06(W)
Lysine	Seed Lys 2-6	Gm 6	C2	Gm06_45433980_G_A	57.2	6.4	52.0-60.3	7.4	0.18 (E)
Lysine	Seed Lys 2-7	Gm 7	M	Gm07_14773717_G_T	48.7	3.4	38.2-87.8	5.6	0.14(W)
Lysine	Seed Lys 2-8	Gm 9	K	Gm09_40970267_C_T	60.0	5.9	56.6-64.9	8.6	0.17(E)
Lysine	Seed Lys 2-9	Gm 13	F	Gm13_35823484-A_G	184.0	8.8	181.2-187.0	12.0	0.20(W)
Lysine	Seed Lys 2-10	Gm19	L	Gm19_42089062_C_T	196.0	4.3	178.0-213.0	5.4	0.14 (W)
Methionine	Seed Met 2-3	Gm 9	K	Gm09_1723633_G_A	0.0	6.0	0.0-4.85	7.4	0.06 (W)
Methionine	Seed Met 2-4	Gm 13	F	Gm13_35823484_A_G	184.0	5.4	183.0-191.9	6.9	0.06(W)
Methionine	Seed Met 2-5	Gm 18	G	Gm18_2020495_C_T	5.1	3.1	5.0-12.0	4.0	0.05 (W)
Threonine	Seed Thr 2-4	Gm 6	C2	Gm06_45871481_C_T	58.0	4.8	51.0-83.0	5.7	0.10 (E)
Threonine	Seed Thr 2-5	Gm 7	M	Gm07_9913651_T_C	43.3	4.9	41.7-69.6	6.9	0.09 (W)
Threonine	Seed Thr 2-6	Gm 9	K	Gm09_38958410_A_G	64.1	5.4	59.0-66.0	8.3	0.10(E)
Threonine	Seed Thr 2-7	Gm 13	F	Gm13_35370448_C_T	183.0	10.5	181.2-187.0	14.3	0.14 (W)
Tryptophan	Seed Tryp 2-2	Gm 6	C2	Gm06_44116624_T_C	54.1	5.6	51.0-60.0	7.3	0.05(E)
Tryptophan	Seed Tryp 2-3	Gm 13	F	Gm13_36316916_C_T	188.3	5.3	182.0-204.6	5.0	0.03 (W)
Tryptophan	Seed Tryp 2-4	Gm 19	L	Gm19_45101232_G_A	201.4	3.8	194.5-229.3	3.8	0.03 (W)

† Chr., chromosome.
‡ MLG=Molecular Linkage Group
§ The QTL position was determined based on genetic linkage map constructed in the present study, measured in centimorgans.
¶ LOD, logarithm of the odds.
# Effect indicates quantitative change in amino acid composition associated with either (E) Essex 15-86-1 or (W) Williams 82-11-43-1 allele.

Table 3.6 Quantitative trait loci identified using R/qtl for composite interval mapping located on various chromosomes associated with amino acid composition in 302 $F_{5:11}$-derived recombinant inbred lines of Essex 86-15-1 x Williams 82-11-43-1 grown in 2013 in Knoxville, TN; Springfield, TN; and Milan, TN.

Location	Trait	QTL Name	Chr†	MLG‡	Molecular Marker	Loc § (cM)	LOD ¶	Confidence Interval of QTL position	R^2 (%)	Effect # g kg^{-1}
Springfield, TN	Cysteine	Seed Cys 2-6	Gm 9	K	Gm09_1768049_G_A	0.2	4.3	0.1-4.9	6.3	0.10(W)
Knoxville, TN	Cysteine	Seed Cys 2-7	Gm 13	F	Gm13_38249824_T_C	199.0	4.2	195.0-201.9	5.0	0.11(W)
Knoxville, TN	Lysine	Seed Lys 2-11	Gm 6	C2	Gm06_44869374_T_G	56.0	3.6	51.0-66.0	4.2	0.19 (E)
Milan, TN	Lysine	Seed Lys 2-12	Gm 6	C2	Gm06_47758592_C_T	65.5	3.9	0.0-69.0	5.8	0.15(E)
Milan, TN	Lysine	Seed Lys 2-13	Gm 7	M	Gm07_18237983_G_A	66.3	4.0	39.2-75.4	6.3	0.14(W)
Knoxville, TN	Lysine	Seed Lys 2-14	Gm 9	K	Gm09_38516865_G_A	61.6	5.8	3.0-64.6	7.2	0.20(E)
Milan, TN	Lysine	Seed Lys 2-15	Gm 9	K	Gm09_37902095_A_G	58.0	3.8	3.0-65.6	5.1	0.13 (E)
Knoxville, TN	Lysine	Seed Lys 2-16	Gm 13	F	Gm13_36031702_T_C	186.6	7.9	181.2-191.0	10.2	0.24(W)
Milan, TN	Lysine	Seed Lys 2-17	Gm 13	F	Gm13_35823484_A_G	185.0	4.8	183.0-191.0	6.5	0.15(W)
Knoxville, TN	Lysine	Seed Lys 2-18	Gm 19	L	Gm19_41845329_T_G	194.8	3.3	186.9-201.6	3.4	0.14(W)
Springfield, TN	Methionine	Seed Met 2-6	Gm 5	A1	Gm05_36780019_T_G	75.7	3.0	64.4-112.7	4.8	0.06 (E)
Knoxville, TN	Methionine	Seed Met 2-7	Gm 9	K	Gm09_1888876_A_G	2.0	4.9	0.0-77.4	5.9	0.08(W)
Springfield, TN	Methionine	Seed Met 2-8	Gm 9	K	Gm09_38956335_G_A	64.1	3.7	0.2-66.5	4.5	0.06 (E)
Milan, TN	Methionine	Seed Met 2-9	Gm 9	K	Gm09_2784336_T_G	6.1	3.5	1.0-64.6	4.2	0.06(W)
Knoxville, TN	Methionine	Seed Met 2-10	Gm 13	F	Gm13_3582384_A_G	185.0	3.7	184.0-194.57	5.1	0.07(W)
Milan, TN	Methionine	Seed Met 2-11	Gm 18	G	Gm18_2007638_G_T	4.7	4.7	0.8-8.0	6.4	0.07(W)
Springfield, TN	Threonine	Seed Thr 2-8	Gm 2	D1b	Gm_02_11030750_C_T	24.0	3.5	9.67-38.7	4.0	0.08(W)
Springfield, TN	Threonine	Seed Thr 2-9	Gm 6	C2	Gm06_44869374_T_G	56.0	3.7	46.0-81.0	4.4	0.11(E)

Table 3.6 (Continued)

Location	Trait	QTL Name	Chr†	MLG‡	Molecular Marker	Loc § (cM)	LOD ¶	Confidence Interval of QTL position	R^2 (%)	Effect # g kg^{-1}
Milan, TN	Threonine	Seed Thr 2-10	Gm 6	C2	Gm06_47764344_C_T	65.5	4.1	56.6-85.0	6.1	0.10(E)
Knoxville, TN	Threonine	Seed Thr 2-11	Gm 7	M	Gm07_10236359_A_G	43.7	3.4	38.5-48.7	6.1	0.11(W)
Milan, TN	Threonine	Seed Thr 2-12	Gm 7	M	Gm07_18237983_G_A	66.3	3.8	62.5-72.0	6.0	0.09(W)
Knoxville, TN	Threonine	Seed Thr 2-13	Gm 9	K	Gm09_38637679_A_C	63.0	4.1	56.2-67.0	4.6	0.10 (E)
Springfield, TN	Threonine	Seed Thr 2-14	Gm 9	K	Gm09_38385411_A_G	61.1	4.7	56.6-66.4	6.7	0.12 (E)
Milan, TN	Threonine	Seed Thr 2-15	Gm 9	K	Gm09_38887894_C_T	63.9	4.0	59.0-66.0	4.6	0.08(E)
Knoxville, TN	Threonine	Seed Thr 2-16	Gm 13	F	Gm13_35823484_A_G	184.0	7.2	183.0-191.0	11.0	0.15(W)
Springfield, TN	Threonine	Seed Thr 2-17	Gm 13	F	Gm13_35823484_A_G	182.0	7.6	180.2-187.0	6.7	0.14 (W)
Milan, TN	Threonine	Seed Thr 2-18	Gm 13	F	Gm13_36573410_T_G	189.0	5.1	183.9-191.0	8.0	0.12 (W)
Milan, TN	Threonine	Seed Thr 2-19	Gm 19	L	Gm19_45158221_T_C	201.6	4.3	199.1-207.0	4.7	0.08 (E)
Springfield, TN	Tryptophan	Seed Tryp 2-5	Gm 6	C2	Gm06_45871481_C_T	58.0	5.8	46.0-58.2	4.9	0.05 (E)
Milan, TN	Tryptophan	Seed Tryp 2-6	Gm 6	C2	Gm06_44150819_G_T	55.0	5.3	50.7-58.3	6.8	0.07 (E)
Knoxville, TN	Tryptophan	Seed Tryp 2-7	Gm 9	K	Gm09_1888876_A_G	2.0	18.1	0.0-3.97	19.2	0.08 (W)
Springfield, TN	Tryptophan	Seed Tryp 2-8	Gm 9	K	Gm09_38013391_A_G	59.0	6.5	47.0-63.0	5.8	0.05 (E)
Milan, TN	Tryptophan	Seed Tryp 2-9	Gm 9	K	Gm09_1888876_A_G	2.0	17.1	0.0-4.85	19.0	0.10(W)
Milan, TN	Tryptophan	Seed Tryp 2-10	Gm 10	O	Gm10_1623075_C_T	2.5	3.5	0.0-14.3	2.0	0.03(W)
Knoxville, TN	Tryptophan	Seed Tryp 2-11	Gm 13	F	Gm13_3301099_T_C	15.8	3.8	0.6-204.6	4.0	0.04 (E)
Springfield, TN	Tryptophan	Seed Tryp 2-15	Gm 13	F	Gm13_35370448_C_T	183.0	8.5	199.0-204.6	3.1	0.04(W)
Milan, TN	Tryptophan	Seed Tryp 2-16	Gm 13	F	Gm13_39252904_A_G	204.3	4.9	0.0-51.1	3.1	0.04(W)
Knoxville, TN	Tryptophan	Seed Tryp 2-17	Gm 18	G	Gm18_2007638_G_T	4.7	3.2	0.0-8.1	2.5	0.02 (W)
Springfield, TN	Tryptophan	Seed Tryp 2-18	Gm 19	L	Gm19_42116080_C_T	196.6	3.3	195.0-219.6	3.4	0.04(W)
Milan, TN	Tryptophan	Seed Tryp 2-19	Gm 19	L	Gm19_42143190_T_C	196.7	7.1	194.1-202.2	5.7	0.05(W)

† Chr., chromosome.
‡ MLG=Molecular Linkage Group
§ The QTL position was determined based on genetic linkage map constructed in the present study, measured in centimorgans.
¶ LOD, logarithm of the odds.
# Effect indicates quantitative change in amino acid composition associated with either (E) Essex 15-86-1 or (W) Williams 82-11-43-1 allele.

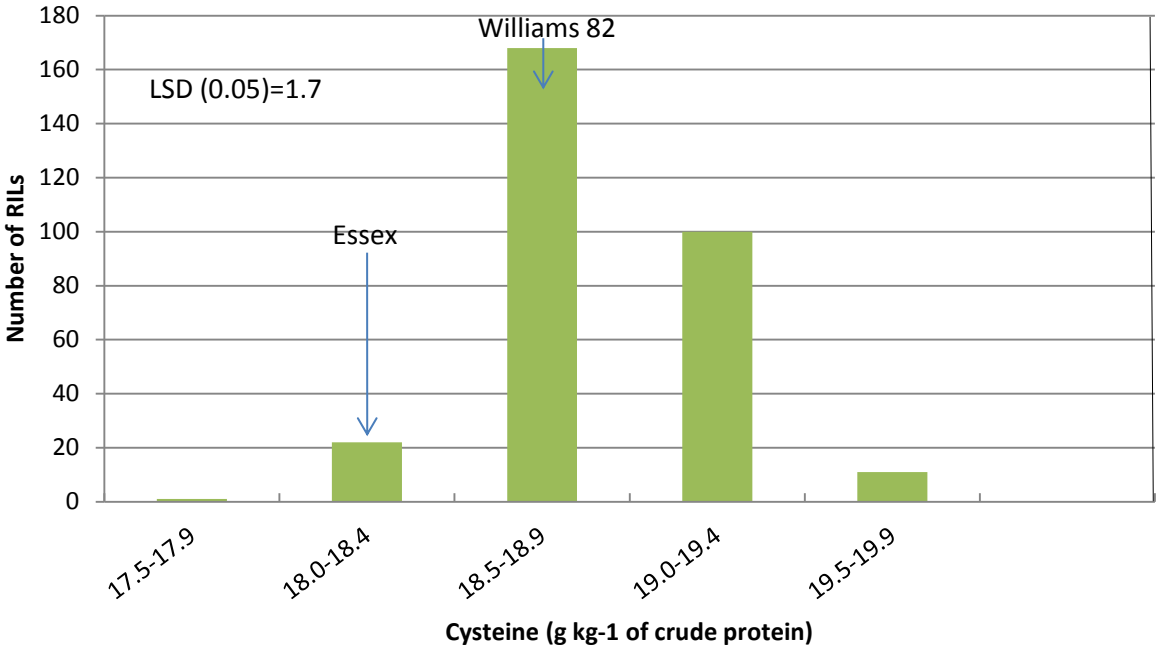

Figure 3.1 Frequency distribution of cysteine for 302 $F_{5:11}$ recombinant inbred lines of Essex *x* Williams 82 averaged over three environments.

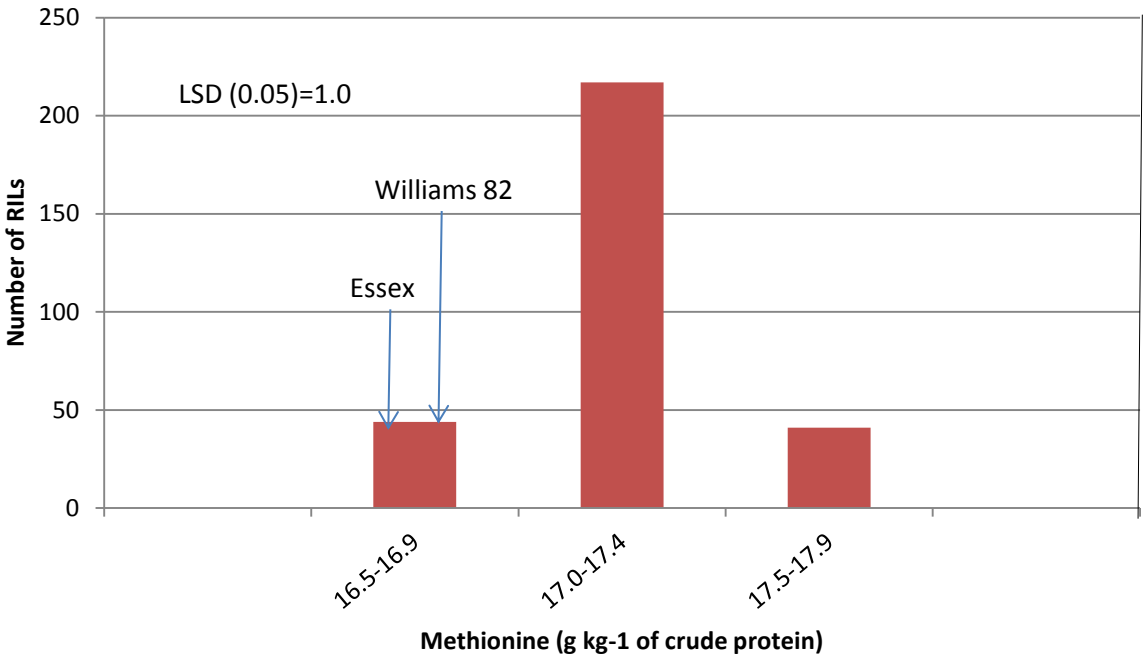

Figure 3.2 Frequency distribution of methionine for 302 $F_{5:11}$ recombinant inbred lines of Essex *x* Williams 82 averaged over three environments.

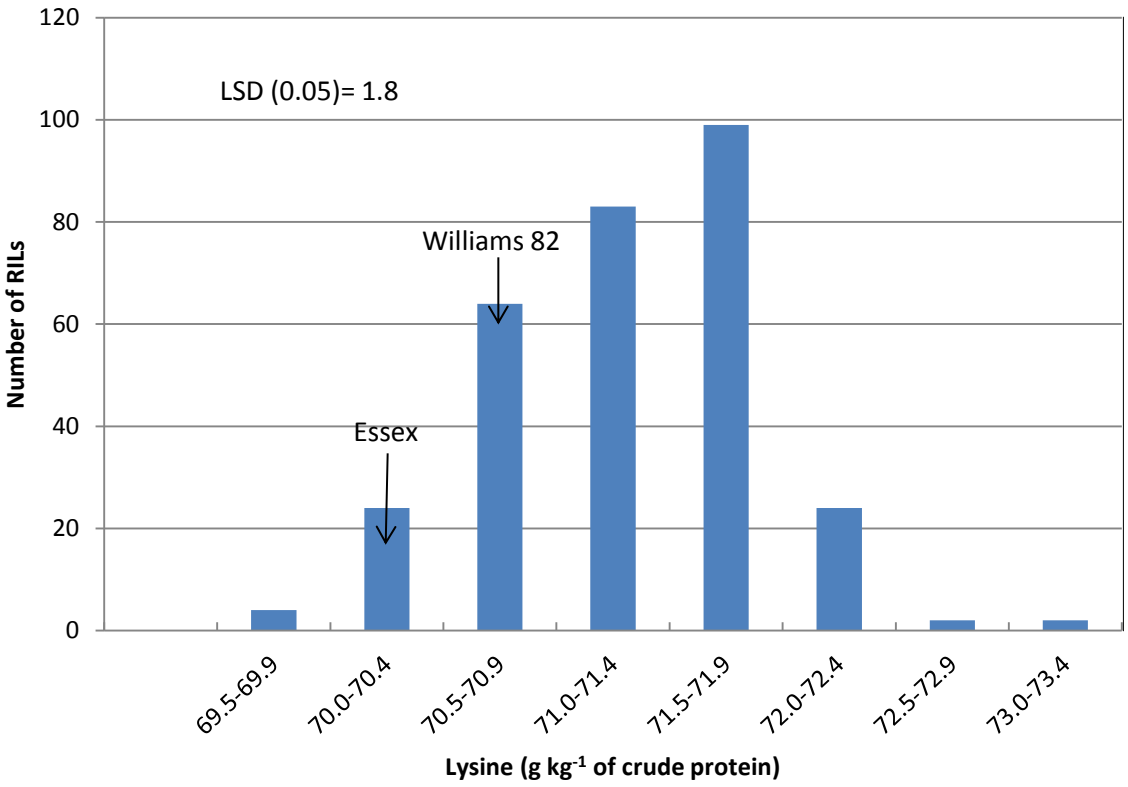

Figure 3.3 Frequency distribution of lysine for 302 $F_{5:11}$ recombinant inbred lines of Essex *x* Williams 82 averaged over three environments.

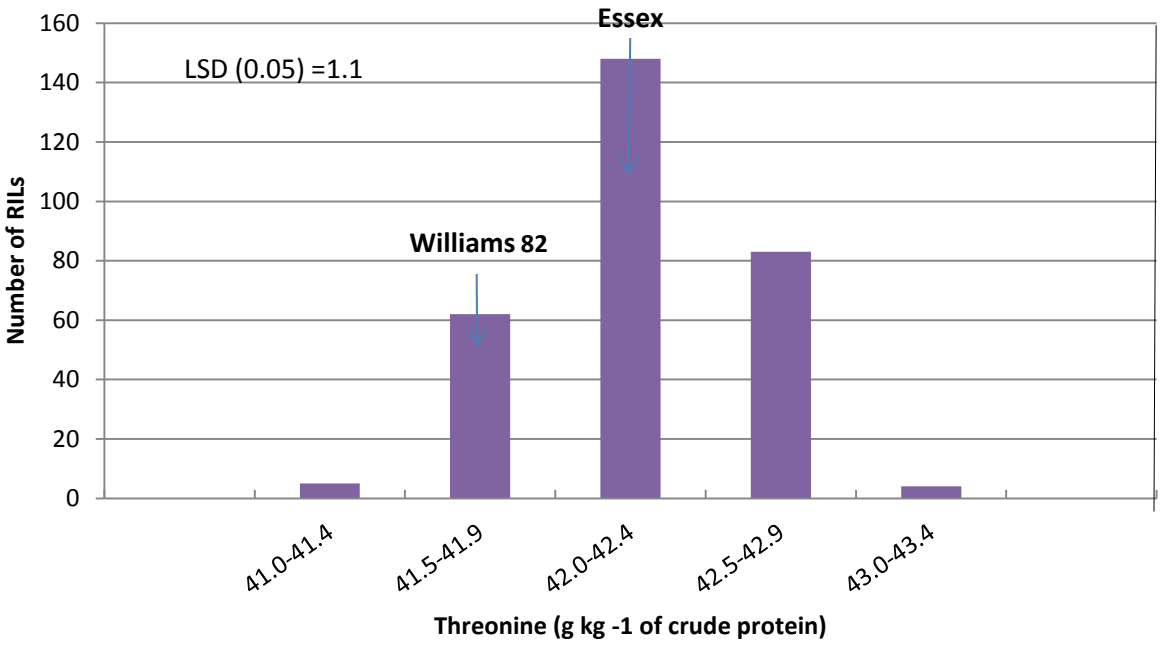

Figure 3.4 Frequency distribution of threonine for 302 $F_{5:11}$ recombinant inbred lines of Essex x Williams 82 averaged over three environments.

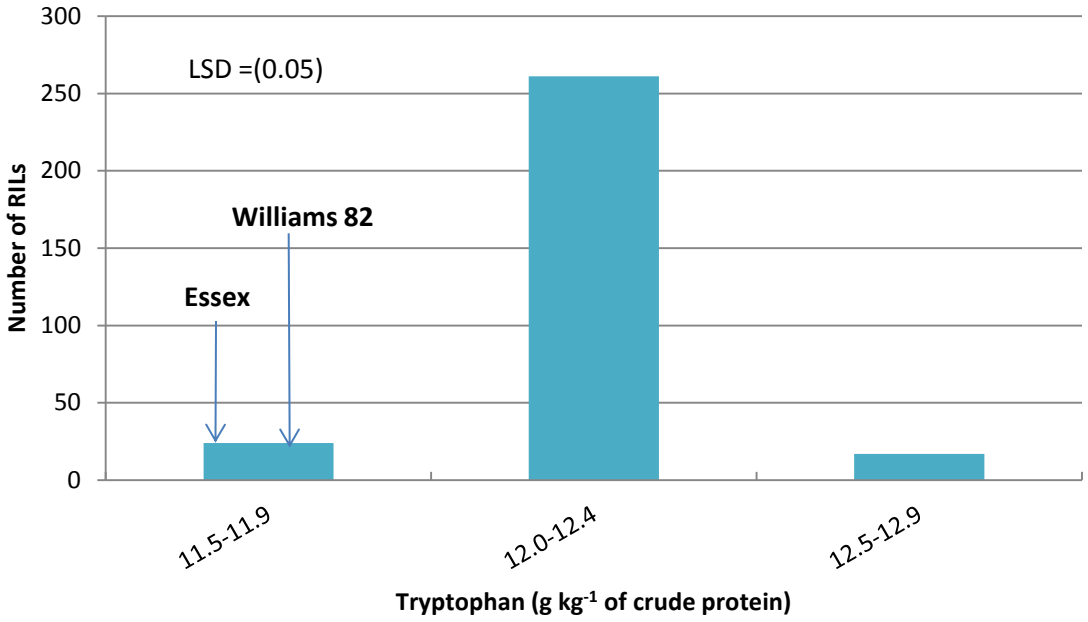

Figure 3.5 Frequency distribution of tryptophan for 302 $F_{5:11}$ recombinant inbred lines of Essex x Williams 82 averaged over three environments.

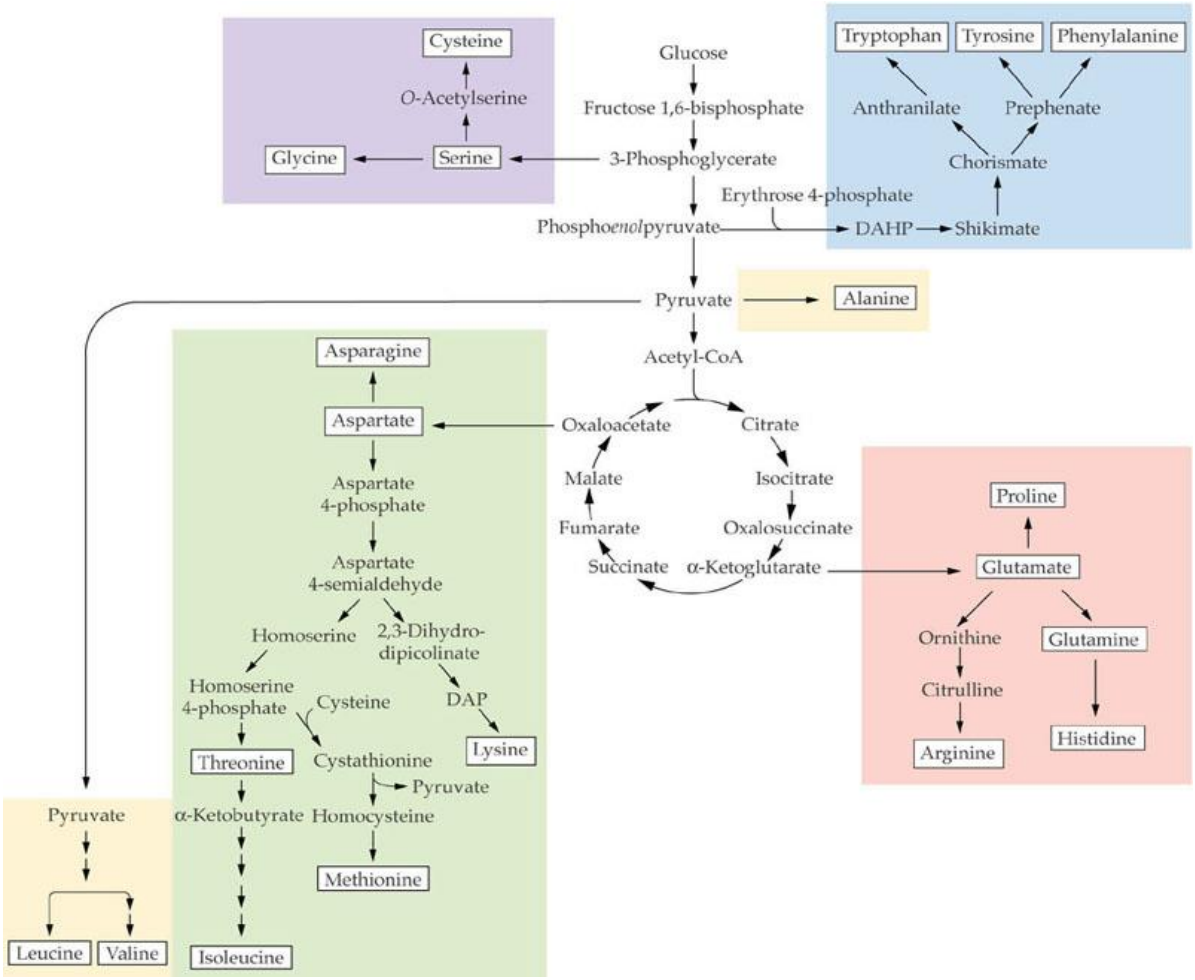

Figure 3.6 A diagram of the pathways responsible for the biosynthesis of the amino acids cysteine, lysine, methionine, threonine and tryptophan (Hildebrand, 2010). Reproduced with the permission of the author.

Chapter 4: Conclusions

Conclusion

The primary goals of this research project were to detect, verify and confirm quantitative trait loci (QTL) for protein concentration, amino acid composition and yield in the Essex *x* Williams 82 population and determine the correlation of amino acids cysteine, methionine, lysine, threonine and tryptophan with major seed quality traits. We successfully identified nine protein QTLs and confirmed a protein QTL on Gm 7 at 50.0 cM. We proposed the name cqSeed Protein-004 for this QTL to the Soybean Genetics Committee. If QTL of this nature are utilized in breeding programs they will enhance the efficiency of marker assisted selection. Nine seed oil QTLs were identified. Yield QTLs were not detected in this experiment. Yield is a complex quantitative trait, we recommend a higher number of genotypes for a yield QTL study to increase statistical power. In the amino acid analysis, 32 amino acid QTLs were identified. The amino acid QTLs will be a substantial addition to current knowledge because few amino acids QTLs have been detected in soybean in comparison to other traits such as protein and yield. While it is important to continue additional studies to identify QTL associated with seed protein, seed oil, yield and amino acids, we must focus on confirmation of QTL that have been discovered in soybean.

Two types of QTL analyses were conducted in this study: combined and individual. The combined analysis successfully detected several QTLs that were consistent across all environments. The individual location analysis detected QTLs that appeared in some environments and were absent in others. The results show the inconsistency of some QTLs across environments, therefore breeders should rely on phenotypic information and incorporate marker assisted selection with confirmed and verified QTLs to help achieve improvements in seed protein and seed oil concentration and amino acid composition.

Vita

Jeneen Abrams is a native of Philadelphia, PA. She is a graduate of W. B. Saul High School of Agriculture Sciences. She received her bachelor degree from Penn State University in Agriculture Sciences and completed her Masters in Plant and Soil Science at Alabama A & M University majoring in Plant Molecular Genetics. She enrolled in the PhD program at the University of Tennessee in fall 2011. She has earned a PhD in Plant Sciences, majoring in Plant Breeding & Genetics and studied under renown soybean breeder, Dr. Vincent Pantalone. Her research has focused on improving protein and oil concentration and amino acid composition in soybean. She plans to continue her career in industry and government, working to assure that we have a healthy and viable food supply.